ADVANCING PRECISION MEDICINE THROUGH INTEGRATIVE BIOINFORMATICS APPROACHES FOR ROBUST BIOLOGICAL KNOWLEDGE DISCOVERY

A Dissertation
Presented to
The Academic Faculty

by

Po-Yen L. Wu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2017

ADVANCING PRECISION MEDICINE THROUGH INTEGRATIVE BIOINFORMATICS APPROACHES FOR ROBUST BIOLOGICAL KNOWLEDGE DISCOVERY

Approved by:

Dr. May D. Wang, Advisor
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. William T. Mahle
Department of Pediatrics
Children's Healthcare of Atlanta

Dr. Robert J. Butera
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Peng Qiu
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Omer T. Inan
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: January 19, 2017

To My Family

ACKNOWLEDGEMENTS

The work discussed in this dissertation would not have been possible without the advice, support, and encouragement of many others, and I would like to use this space to acknowledge their generous help during my Ph.D. journey.

I would like to first thank Dr. May Wang, my thesis advisor, for supporting me over the last several years and allowing me to pursue my strength as a researcher. Dr. Wang provided not only impactful foresight in my thesis research but also insightful guidance in my career development. She consistently convinced many very important people to listen to and devote resources to the ideas and investigations of her graduate students. I really appreciate the training and guidance I received from her.

I am very thankful to the four professors who served on my Ph.D. thesis committee—Dr. Robert Butera, Dr. Omer Inan, Dr. William Mahle, and Dr. Peng Qiu. They gave me helpful feedback and took time out of their busy schedules to examine and critique my research. Their guidance greatly improved the quality of my work, and I really appreciate every comment they provided.

I am also very grateful to the professors, researchers, and doctors with whom I had the opportunity to collaborate. Dr. Kevin Maher, Dr. Nikhil Chanani, and Dr. Greg Martin in Children's Healthcare of Atlanta and Emory University not only shared their clinical knowledge and experience for our collaborative real-world projects but also provided insightful research directions from physicians' viewpoints. While working on the FDA-led international collaborative project, I received tons of valuable comments and feedback from many collaborators, including Dr. Wendell Jones, Dr. Leming Shi, Dr.

Matthias Fischer, Dr. Christopher Mason, Dr. Joshua Xu, Dr. Wei Shi, Dr. Jian Wang, Dr. Jean Thierry-Mieg, Dr. Danielle Thierry-Mieg, Dr. David Kreil, Dr. Dalila Megherbi, Dr. Gary Schroth, and Dr. Weida Tong. They are all experts in bioinformatics research, and their constructive feedback and suggestion really improved the quality of my work.

Many fellow researchers and students in the Biomedical Informatics and Bioimaging Laboratory (Bio-MIBLab) contributed in various ways to my work. Dr. John Phan deserves my special thanks for mentoring me through the first several years of my Ph.D. I learned many things from him, including, but not limited to, the thinking process towards challenges, technical writing skills, communication skills, and most importantly, working styles. Dr. R. Mitchell Parry, Dr. Todd Stokes, and Dr. Richard Moffitt were post-docs in my lab, and they helped me improve my work by critically critiqued my research while I was writing papers or presenting in group meetings. Fellow Ph.D. students, Dr. Sonal Kothari Phan, Dr. Chih-Wen Cheng, Dr. Chanchala Kaddi, Janani Venugopalan, Ryan Hoffman, Li Tong, Hamid Hassanzadeh, Ying Sha, and Hang Wu were excellent companions along my Ph.D. journey. Their support, either technically or morally, is the most invaluable asset that I will cherish forever.

Finally, I would like to thank my parents, my younger sister, and all my friends for their endless support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

AdaBoost	Adaptive Boosting
ANOVA	Analysis of Variance
AUC or AUROC	Area Under the Receiver Operating Characteristic Curve
BGI	Beijing Genomics Institute
BP	Base Pairs
CPB	Cardiopulmonary Bypass
CCDS	Consensus Coding Sequences
cDNA	Complementary Deoxyribonucleic Acid
ChIP-seq	Chromatin Immunoprecipitation Sequencing
CNV	Copy Number Variation
CoV	Coefficient of Variation
Ct	Cycle Threshold
CVD	Cardiovascular Disease
DEG	Differentially Expressed Gene
DIKW	Data, Information, Knowledge, and Wisdom
DNA	Deoxyribonucleic Acid
EFS	Event-Free Survival
EHR	Electronic Health Record
EM	Expectation-Maximization Algorithm
EMR	Expected Mixing Ratio
ERCC	External RNA Controls Consortium
FC	Fold Change

FIMO	Find Individual Motif Occurrences
FPKM	Fragments per Kilobase per Million Mapped Fragments
FPM	Fragments per Million Mapped Fragments
GO	Gene Ontology
GTF	Gene Transfer Format
GWAS	Genome-Wide Association Study
HBRR	Human Brain Reference RNA
ICC	Intraclass Correlation Coefficient
ICU	Intensive Care Unit
INDEL	Insertion Mutation and Deletion Mutation
INSDC	International Nucleotide Sequence Database Collaboration
iPOP	Integrative Personal Omics Profile
LOS	Length of Stay
LR	Logistic Regression
LQ	Lower Quartile
MAD	Mean Absolute Deviation
MAQC	Microarray Quality Control
MAY	Mayo Clinic
MBD-seq	Methyl-CpG-Binding Domain Sequencing
MCC	Matthews correlation coefficient
mRMR	Minimum Redundancy, Maximum Relevance
MT	Mitochondria
NGS	Next-Generation Sequencing
OS	Overall Survival
P4 Medicine	Predictive, Personalized, Preventive, and Participatory Medicine

qPCR	Quantitative Polymerase Chain Reaction
RLE	Relative Log Expression
RMSE	Root-Mean-Squared Error
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
RPKM	Reads per Kilobase per Million Mapped Reads
RPM	Reads per Million Mapped Reads
rRNA	Ribosomal RNA
SAM	Sequence Alignment/Map
SD	Standard Deviation
SEQC	Sequencing Quality Control
SNP	Single Nucleotide Polymorphism
SRA	Sequence Read Archive
SV	Structural Variation
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas
TMM	Trimmed Mean of M-Values
TO	Titration Order
TPM	Transcripts per Million
UCSC	University of California, Santa Cruz
UHRR	Universal Human Reference RNA
UQ	Upper Quartile

SUMMARY

Facilitated by -omic data, precision medicine is a promising medical model that may revolutionize the quality of the current healthcare system. Currently, -omic data are being rapidly accumulated because of the advent of high-throughput -omic assays. Though challenging, abundant information embedded in these data is encouraging for the realization of precision medicine. Data analytics, including data pre-processing and data modeling techniques, has been successfully applied to many -omic applications, and biomarkers identified from -omic data are viewed as catalyzers for precision medicine.

The goal of my Ph.D. research is to address some key challenges in the process from raw -omic data to disease subgroup assignment for precision medicine, including (1) the lack of standardized bioinformatics pipelines that extract high-quality gene expression from the raw RNA sequencing data; (2) the lack of systematic, quantitative assessment of the contribution of upstream bioinformatics pipeline components to downstream variations in identified biomarkers or clinical endpoint prediction performance; and (3) the lack of effective strategies for integrating knowledge derived from multiple -omic data sources, either the same type or different types. This dissertation addresses these challenges through three specific aims:

(1) Quality Control for Precision Medicine: to investigate and control the impact of bioinformatics pipelines on feature quality using RNA sequencing data.

(2) Knowledge Discovery for Precision Medicine: to discover impactful biomarkers that facilitate disease subgroup assignment using NGS data

(3) Integrative Analysis for Precision Medicine: to integrate multiple sources of -omic data for improved disease subgroup assignment.

The research in this dissertation was completed in frequent collaboration with the Food and Drug Administration, Children's Healthcare of Atlanta, Emory University, and Georgia Institute of Technology. Proposed analytical approaches for NGS data have been systematically evaluated and validated using a variety of experimental designs with various NGS datasets. These results and associated case studies demonstrate the contribution of this work to and its future potential in the paradigm shift from current pattern-based, evidence-based medicine to future algorithm-based precision medicine.

CHAPTER 1

INTRODUCTION

Providing the best health care for patients is always the driving force for the transformation of the healthcare system. Many medical models have been proposed for improving the healthcare system, and personalized medicine was the most attractive, promising model, aiming to customize healthcare delivery for each individual and to maximize the effectiveness of each patient's treatment [1]. The concept of personalized medicine has had a long history, but its development really speeded up only after the completion of the Human Genome Project, leading to the rapid advancement of functional genomics that in turn enables more precise risk estimation and therapeutic response prediction [2]. A further improvement of the current healthcare system was advocated by Hood *et al.* since 2009. They proposed the concept of predictive, personalized, preventive, and participatory medicine (P4 medicine), which aims to transform current reactive medicine to future proactive medicine [3]. Such a major paradigm shift may potentially reduce healthcare expenditure and ameliorate patients' prognosis. Recently, momentum has been accumulated for moving from personalized medicine to precision medicine. The subtle but fundamental difference between personalized medicine and precision medicine is that precision medicine pursues a more precise classification of patients into subgroups that share a common biological basis of diseases [4, 5]. Such a precise classification may potentially lead to more effective treatments and better clinical outcomes, and this is exactly the promise of precision medicine advocated by President Barack Obama in his State of the Union Address in

January 2015: "… delivering the right treatments, at the right time, every time to the right person …" [6].

Precision medicine presents a promising paradigm shift from early symptom-based intuition medicine, today's patterned-based, evidence-based medicine, to future algorithm-based medicine (**Figure 1**). The key to precision medicine centers around the data, including data collection, data management (e.g., data storage, data sharing, and data privacy), and data analytics (e.g., data mining, data integration, data interpretation, and data visualization) [7]. All these data-related components were also brought up in a statement from The White House: "… the Precision Medicine Initiative will leverage advances in genomics, emerging methods for managing and analyzing large data sets …" [8]. As technologies evolved, a tremendous amount of biomedical data has been generated and stored. These biomedical data are viewed as big data because of their high complexity and typically large volume [9]. Currently, biomedical big data and associated big data analytics have been applied to several key research areas, including, but not limited to, bioinformatics [10-12], health informatics [13-16], biomedical imaging informatics [17-19], and biosensor informatics [20-22]. It was only recently that -omic data were identified as the main catalyzer for precision medicine, and integrating -omic data into the current electronic health record (EHR) system is one most probable approach for the realization of precision medicine [4, 7, 23, 24].

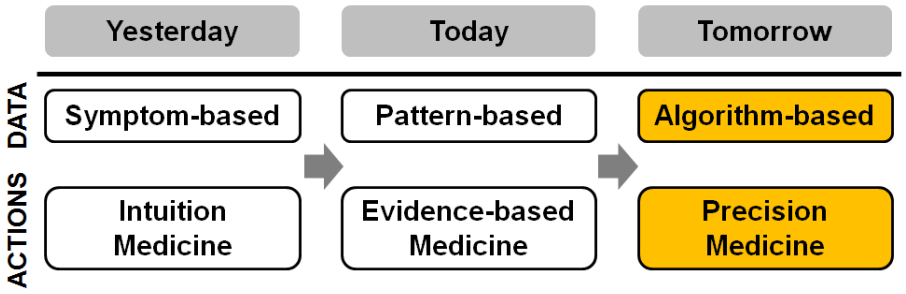


Figure 1: Evolution of Healthcare Models. Models for health care are ever-changing, and their ultimate goals are to provide the most effective treatment for every patient. With tons of stored biomedical big data and the advancement in algorithms and analytics, it is promising to move from today's pattern-based, evidence-based medicine to future algorithm-based precision medicine.

The fast accumulation of -omic data largely owes to the invention of several high-throughput technologies such as next-generation sequencing (NGS), microarrays, and high-resolution mass spectrometry (**Figure 2**). These high-throughput technologies facilitate the collection of various types of -omic data, including genomic, transcriptomic, epigenomic, proteomic, and metabolomic (**Figure 2**). Each type of -omic data possesses its unique aspects of personalized information about a patient. Thus, by integrating knowledge derived from various types of -omic data, it is possible to identify a set of rules or characteristics that may help precisely categorize patients into disease subgroups, which is the essence of precision medicine. The concept of precision medicine is promising, and the data that support this concept has been constantly generated. However, the connection between the concept and the raw -omic data is still in its infancy. With such overwhelming data complexity and volume, it is necessary to have novel analytics for handling and analyzing these big data (**Figure 2**) [25]. Following the DIKW (data, information, knowledge, and wisdom) hierarchy [26], -omic data analytics mainly focuses on extracting molecular profiles from raw data generated by high-

throughput assays, deriving biological knowledge from raw molecular information, integrating knowledge derived from various types of -omic data, and finally concluding with actionable knowledge (i.e., wisdom) for different clinical applications (**Figure 3**).

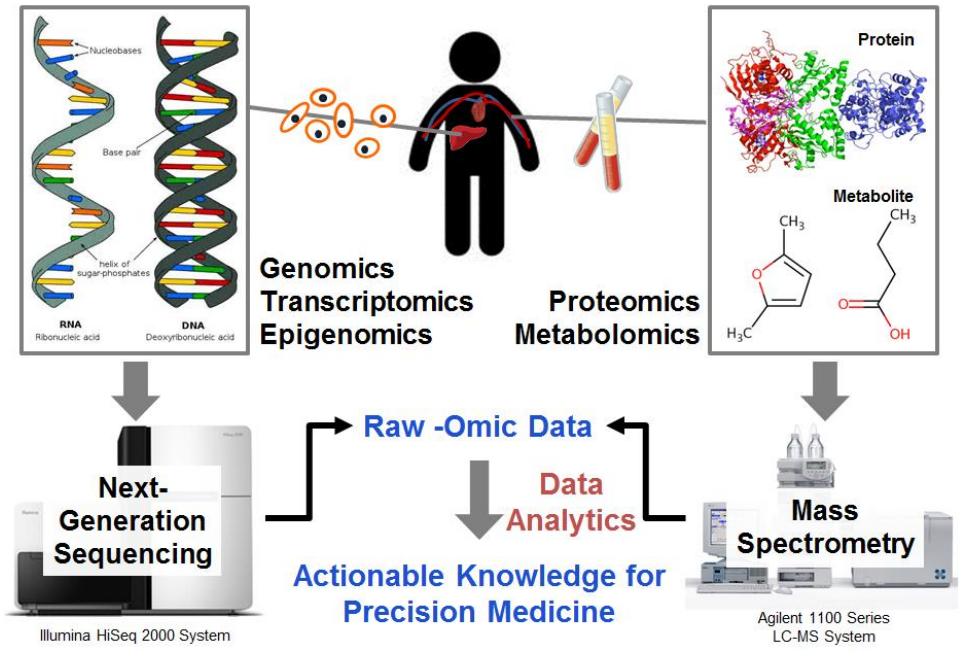


Figure 2: Personal Molecular Fingerprints. Each individual has unique molecular fingerprints that are mainly composed of DNA, RNA, protein, and metabolite profiles. Next-generation sequencing and mass spectrometry are two popular high-throughput assays that help extracting raw -omic data, including genomic (DNA), transcriptomic (RNA), epigenomic (DNA), proteomic (proteins), and metabolomic (metabolites). With proper data analytics, it is possible to identify actionable knowledge for precision medicine using these raw -omic data.

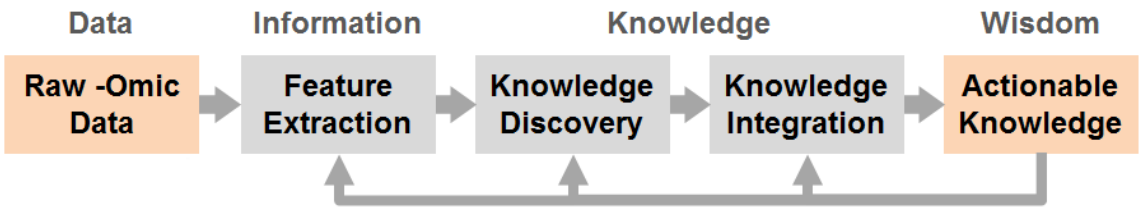


Figure 3: DIKW Hierarchy for -Omic Data. DIKW stands for data, information, knowledge, and wisdom. Following the DIKW hierarchy, data analytics for -omic data plays the role of extracting personal molecular fingerprints (Data to Information), discovering all available -omic knowledge (Information to Knowledge), integrating knowledge from multiple -omic data (Knowledge to Wisdom), and finally identifying actionable knowledge (Wisdom) for precision medicine.

Among high-throughput -omic assays, the most popular one for acquiring genomic, transcriptomic, and epigenomic data is the NGS technology. Therefore, the majority of my thesis work centered on developing data analysis pipelines for NGS data, with a focus on RNA sequencing, which is one of the most important applications of the NGS technology. The rest of this introduction aims to provide some background information about my thesis work, beginning with the NGS technology, followed by various types of -omic data (i.e., genomic, transcriptomic, and epigenomic) generated by NGS and associated -omic features. I then introduce the current state of the art in feature extraction and biomarker identification for NGS-based -omic data, followed by the detailed elaboration of RNA sequencing expression analysis pipelines. Finally, I end with a discussion of bioinformatics challenges in the DIKW hierarchy for NGS data that motivate my thesis work formulated in three specific aims.

1.1 Next-Generation Sequencing Technologies

Commercially available and widely applied NGS platforms include Illumina, Ion Torrent, and Roche/454. These platforms differ primarily in the experimental basis and

protocol but share similar sequencing steps. In general, after sequencing library preparation, DNA/cDNA fragments are isolated and attached to a substrate. These fragments are amplified and then sequenced in parallel. The iterative sequencing steps, followed by fluorescence imaging, produce a large set of imaging data that will be translated into sequence "reads" by platform-specific base-calling algorithms.

Illumina is the most prevalent NGS platform that uses a "sequencing-by-synthesis" technique. DNA/cDNA molecules are fragmented, ligated with adapters, attached to a proprietary flow cell surface, followed by a bridge amplification process that forms clonally distributed clusters for each attached DNA/cDNA fragments. Using a cyclic reversible termination technique, four fluorescently-labeled nucleotides are fed to the flow cell with one and only nucleotide binds to each surface-bound fragment. These fluorescent dyes are then imaged and washed away before the next sequencing cycle [27]. Ion Torrent leveraged Moore's Law in manufacturing semiconductors and invented a semiconductor chip that is capable of directly translating chemical signals into digital information. Similar to Illumina, adapter-ligated DNA/cDNA fragments are attached to the proprietary Ion Sphere particles, followed by emulsion PCR before depositing into wells on a semiconductor chip where the sensing process occurs [28]. Roche/454 pyrosequencing is the earliest commercialized NGS platform also based on emulsion PCR for amplification. It determines nucleotide sequences by detecting the release of pyrophosphates when nucleotides are cyclically added [29].

1.2 NGS-based -Omic Data

-Omic data contain a comprehensive catalog of molecular profiles in biological samples. They have been viewed as the fundamental driving force for precision medicine.

Major types of -omic data include genomic, transcriptomic, epigenomic, proteomic, and metabolomic, and NGS is the most popular assay for generating genomic, transcriptomic, and epigenomic data.

The uniqueness property of each person's genome as well as the closely related transcriptome and epigenome provides a promising opportunity for precision medicine. The advent of several high-throughput technologies, such as NGS and DNA microarrays, has offered the capability of studying the entire genome, transcriptome, and epigenome in a faster and more cost-effective manner. A genome contains a complete set of DNAs in a cell. Knowledge embedded in the genome includes single nucleotide polymorphisms (SNPs), insertions, deletions (indels), copy number variations (CNVs), and several other structural variations (SVs) [30, 31]. A transcriptome contains all kinds of RNAs (e.g., mRNA, tRNA, and rRNA) transcribed from the genome. Knowledge embedded in the transcriptome includes gene expression, transcript expression, gene fusion, and alternative splicing [32, 33]. Lastly, an epigenome contains genome-wide chemical modification or marking of DNA sequences. Such modification does not change DNA sequences but does influence downstream transcriptional mechanisms. Knowledge embedded in the epigenome includes genome-wide protein-DNA binding sites, histone modification patterns, and DNA methylation patterns [34]. Epigenomic data may be the most essential building block for pursuing precision medicine due to its role in affecting other -omic data, such as transcriptomic, proteomic, and metabolomic [35].

1.3 Feature Extraction Techniques for NGS-based -Omic Data

Raw data generated by NGS are short sequence reads, or reads for short, that contain the reading of nucleotide sequences, either DNA or cDNA. These raw NGS-

based -omic data are not directly interpretable, and the required feature extraction steps depend on the type of -omic data. Several selected tools for -omic data feature extraction are summarized in **Table 1**.

NGS is a popular assay for genomic, transcriptomic, and epigenomic studies. Sequence mapping, which identifies not only the origin but also the alignment of each read, is the first step for most NGS applications [36]. It is a computationally intensive process that requires auxiliary data structures (e.g., the hash table [37] and the Burrows-Wheeler transform [38]), multithreading, or in-memory computing [39] for improving computational efficiency.

Genomic studies typically aim to identify genomic variants, either small-scale (e.g., SNPs) or large-scale (e.g., SVs), in the sequenced genome [40]. Small-scale variant detection uses per base differences between reads and the reference genome as the evidence [30, 41]. Large-scale variant detection applies various approaches, including read-pair-based, read-depth-based, split-read-based, and assembly-based [42, 43].

For transcriptomic studies, major applications include expression profiling, fusion gene detection, and alternative splicing detection [32, 33]. Expression profiling is a process that associates mapped reads with genes and corresponding transcripts. The main difference among various quantification methods is the handling of multi-mapped reads. Some methods associate the reads with all loci [44, 45], whereas others probabilistically associate the reads with only a few model-inferred loci [46, 47]. Gene fusion is a rare event that two partial genes combine and form a new gene. Fusion gene detection relies on two main evidence: the spanning read pairs, which indicate that a fusion boundary exists between the two ends, and the split read, which provides more definite evidence for

the location of a fusion boundary [48, 49]. Alternative splicing is a process that includes or excludes certain exons when forming mature mRNAs. Detecting alternative splicing relies on either de novo transcriptome assembly [50-54], or inference from sequence-mapping outputs [47, 55].

Table 1: Selected Tools for -Omic Feature Extraction.

Tool	Assay	-Omic Data	Key Functionality
GMAP [37] BWA [38] STAR [39]	Next-generation sequencing	Genomic, transcriptomic, and epigenomic	Sequence mapping
GATK [30] SAMtools [41]		Genomic	Genomic variant discovery
HTSeq [44] BEDTools [45]		Transcriptomic	Gene expression quantification
RSEM [46] Cufflinks [47]			Gene and transcript expression quantification
defuse [48] TopHat-Fusion [49]			Gene fusion detection
Trans-ABySS [50] Trinity [51] Cufflinks [47] Scripture [55]			Alternative splicing detection and quantification
MACS [56] SISSRs [57]		Epigenomic	ChIP-seq peak calling

GMAP stands for genomic mapping and alignment program; BWA, Burrows-Wheeler aligner; STAR, spliced transcripts alignment to a reference; GATK, genome analysis toolkit; RSEM, RNA-seq by expectation-maximization; Trans-ABySS, transcriptome assembly and analysis pipeline; MACS, model-based analysis of ChIP-seq; and SISSRs, site identification from short sequence reads.

Epigenomic studies focus on identifying putative DNA-binding sites, histone modification patterns, and DNA methylation patterns [34]. Since sample preparation protocols ensure only relevant genomic regions are sequenced, the feature extraction of epigenomic data usually contains the following three steps: building a single profile representing the density of reads along the genome, modeling background noises, and finally determining statistically significant peaks [36].

1.4 Biomarker Identification Techniques for NGS-based -Omic Data

Feature extraction techniques help derive interpretable -omic features from the raw -omic data. In practice, multiple groups of samples are collected, representing different biological conditions (e.g., disease verus non-disease) or different time points (e.g., before versus after a treatment). With such an experimental design, it becomes feasible to identify -omic biomarkers that are discriminatory among groups. Several selected tools for biomarker identification are summarized in **Table 2**.

For genomics, not all variants have significant impact on phenotypic traits. The advent of genome-wide association studies (GWASs) helps assess the degree of association between each variant and a targeted trait [58]. However, GWAS presents some big limitations especially when dealing with complex diseases such as cancer and cardiovascular diseases [59]. Most GWAS tools focus on SNP association [60-62], while only a few can infer CNV or SV association [63, 64]. Most other -omic biomarkers are identified by investigating statistically significant differences among groups, such as differentially expressed genes/transcripts [65, 66], differential alternative splicing [67, 68], differential DNA binding [69], differential histone modification [70], and differential methylation [71]. In general, the abundance of each group is quantified and fitted to

Poisson-based distributions (e.g., the Poisson distribution and the negative binomial distribution), followed by statistical tests (e.g., the Fisher's exact test and the likelihood ratio test) that determine the statistical significance of each molecular feature.

Table 2: Selected Tools for -Omic Biomarker Identification.

Tool	-Omic Data	-Omic Biomarker	Approach
SNPassoc [60]	Genomic	Significant SNPs associated with traits	Genome-wide association studies
SNPTEST [61]			
VAT [62]		Significant SNPs and indels associated with traits	
PLINK [63]		Significant SNPs, indels, and CNVs associated with traits	
CNVRuler [64]		Significant CNVs associated with traits	
edgeR [65]	Transcriptomic	Differentially expressed genes /transcripts	Differential analysis (model fitting and statistical tests)
DESeq2 [66]			
DiffSplice [67]		Differential alternative splicing	
MATS [68]			
DBChIP [69]	Epigenomic	Differential binding sites	
ChIPDiff [70]		Differential histone modification sites	
QDMR [71]		Differentially methylated regions	

SNPassoc stands for SNP-based whole genome association studies; VAT, variant association tools; PLINK, population-based linkage analyses; edgeR, empirical analysis of digital gene expression data in R; MATS, multivariate analysis of transcript splicing; DBChIP, differential binding with ChIP-seq data; and QDMR, quantitative differentially methylated regions.

1.5 RNA Sequencing Expression Analysis Pipelines

RNA sequencing, or RNA-seq for short, is a major branch of NGS. It is capable of capturing comprehensive transcriptomic information such as gene expression, transcript expression, gene fusion, and alternative splicing [32]. To capture the highly dynamic transcriptome, one popular application of RNA-seq is to study gene and transcript expression among various biological conditions or samples collected at various

time points [33]. In the first step of RNA-seq expression analysis pipeline, sequence reads are mapped to a reference genome or transcriptome. Next, the quantification step estimates gene or transcript expression. Finally, the normalization step enables inter- or intra-sample comparison. RNA-seq-based inferences typically based on modeling the normalized gene or transcript expression [72] that will be discussed in later chapters.

1.5.1 Sequence Mapping

The first step of bioinformatics pipelines for most NGS-based applications is sequence mapping, and the RNA-seq expression analysis pipeline falls within this category. Sequence mapping determines the genomic or transcriptomic origin of sequence reads (or "reads" for short). The brute-force strategy for sequence mapping requires large CPU and memory resources. For example, mapping millions of reads to the three billion base pairs (bp) of the human genome is extremely time-consuming. Thus, the research of sequencing mapping largely focuses on improving computational efficiency while maintaining high mapping accuracy.

Table 3 lists some mapping tools with their mapping strategies and key features. Depending on biological applications and computational resources, mapping algorithms can provide three types of alignments: (1) Un-gapped alignment (e.g., Bowtie [73]) allows only mismatches between query reads and the reference genome to keep the computational cost low. However, for some applications (e.g., mapping of RNA-seq data to the human genome), un-gapped alignment may fail to align a large number of reads. (2) Gapped alignment (e.g., BFAST [74], Bowtie2 [75], BWA [38], Novoalign [76], SHRiMP2 [77], SOAPaligner [78], and SSAHA2 [79]) allows mismatches, insertions, and deletions. Most gapped alignment tools implement the Smith-Waterman [80] or

Needleman-Wunsch [81] algorithms. (3) Spliced alignment (e.g., GSNAP [82], TopHat [83], MapSplice [84], OSA [85], and SOAPsplice [86]) allows the long extension of gaps within the query reads. Biologically, such long gaps may represent intronic regions or inter-chromosomal splitting. Algorithmically, spliced alignment may be achieved by segmenting query reads into smaller sequences (e.g., 25 bp), mapping these smaller sequences, and then assembling mapped results for each read into a consensus result. Spliced alignment algorithms are often computationally more expensive than un-spliced algorithms. However, spliced mapping is necessary for applications that focus on identifying novel splice junctions using RNA-seq. Un-spliced mapping, including gapped and un-gapped, is sufficient for ChIP-seq data analysis.

Mapping accuracy depends on mapping strategy. Uniquely mapped reads provide more definite information than multi-mapped reads. If a query read is mapped to multiple genomic loci due to short read length, the ambiguous mapping happens and a mapping tool may randomly report one mapping out of all optimal mappings or report all optimal mappings. On the other hand, multi-mapped reads may benefit the downstream quantification algorithms in model training and expression estimation.

To improve the computational efficiency of sequence mapping, auxiliary data structures can be used to reduce the similarity search space such as to index either the reference genome or query reads using hash tables (e.g., BFAST, GSNAP, SHRiMP2, and SSAHA2 are representatives), or to index the reference genome using the Burrows-Wheeler transform with suffix/prefix arrays (e.g., Bowtie, Bowtie2, BWA, and SOAP2 are representatives) [87].

Table 3: Summary of Sequence Mapping Tools for RNA-seq.

Sequence Mapping Tool	Mapping Strategy and Usage	Algorithmic Notes
BFAST [74]	Un-spliced mapping to transcriptome or genome	Hash table, Smith-Waterman local alignment
Bowtie [73]		Burrows-Wheeler transform and FM-index
Bowtie2 [75]		Burrows-Wheeler transform, FM-index-assisted seed alignment, dynamic programming
BWA [38]		Burrows-Wheeler transform
Novoalign [76]		*Commercial software, algorithm un-published
SHRiMP2 [77]		Multiple spaced-seed indexing, Smith-Waterman local alignment
SOAPaligner [78]		Bi-directional Burrows-Wheeler transform
SSAHA2 [79]		Hash table
GSNAP [82]	Spliced mapping to genome & un-spliced mapping to transcriptome or genome	Minimal sampling strategy, oligomer chaining for approximate alignment, sandwich dynamic programming
MapSplice [84]	Spliced mapping to genome	Uses Bowtie for alignment, segmented mapping
OSA [85]		Two-stage transcriptome and genome alignment, segmented mapping
SOAPsplice [86]		Burrows-Wheeler transform, segmented mapping
TopHat [83]		Uses Bowtie or Bowtie2 for alignment, segmented mapping

BFAST stands for BLAT-like fast accurate search tool; BWA, Burrows-Wheeler aligner; SHRiMP2, short read mapping package, version 2; SOAPaligner, short oligonucleotide analysis package aligner; SSAHA2, sequence search and alignment by hashing algorithm, version 2; GSNAP, genomic short-read nucleotide alignment program; OSA, Omicsoft sequence aligner; and SOAPsplice, short oligonucleotide analysis package for splice junction detection

1.5.2 Expression Quantification

The second step of the RNA-seq bioinformatics pipeline is expression quantification of genes or transcripts. Because a read may map to multiple genomic loci, the accuracy of gene or transcript expression estimation depends on the ability of a quantification algorithm to resolve the ambiguities from the sequence-mapping step. In addition, a gene may have multiple alternatively spliced transcripts sharing a common set of exons, where a read mapped to the shared exons may belong to any one of the transcripts. Currently, the handling of these ambiguities involves building a probabilistic framework and then estimating gene or transcript expression using either the expectation-maximization (EM) algorithm or Bayesian inference [46, 47, 88].

Quantification algorithms can be categorized into three groups: count-based, linear model-based, and Poisson model-based [89]. **Table 4** lists common RNA-seq quantification tools, categorized in terms of the model, the estimation algorithm, and quantifiable targets. Count-based quantifiers (e.g., ERANGE [90], HTSeq [44], NEUMA [91], and ALEXA-Seq [92]) assign each read to its mapped location with a probability of one. Each count-based quantifier implements a proprietary filtering criterion, and the expression profile is the accumulated read count on each targeted gene or transcript. Linear model-based quantifiers (e.g., rQuant [93] and IsoInfer [94]) assume that read counts are normally distributed, and least squares can be applied to infer expression estimates. Poisson model-based quantifiers (e.g., RSEM [46], Cufflinks [47], MISO [88], and IsoEM [95]) probabilistically assign multi-mapped reads based on the assumption that reads from genomic loci follow the Poisson distribution. Count-based quantifiers do not rely on a predefined model, so they have lower computational complexity than

model-based quantifiers. However, the expression estimates of count-based quantifiers might deviate from the truth because of the naïve way multi-mapped reads are handled.

Table 4: Summary of Expression Quantification Tools for RNA-seq.

Quantification Tool	Mathematical Model	Estimation	Gene / Transcript
ALEXA-Seq [92]	Count-based model	Average coverage of mapped reads	Yes / Yes
ERANGE [90]		Accumulated counts, read assigns proportionally to expression level	Yes / No
HTSeq [44]		Accumulated counts, read assigns with probability 1	Yes / No
NEUMA [91]		Accumulated counts of informative reads	Yes / Yes
IsoInfer [94]	Linear model	Maximum likelihood estimation from convex quadratic programming	Yes / Yes
rQuant [93]		Minimize read coverage deviation with quadratic programming	Yes / Yes
Cufflinks [47]	Poisson model	Maximize likelihood with the maximum a posteriori estimates using Bayesian inference	Yes / Yes
IsoEM [95]		Maximum likelihood estimation with EM algorithm	Yes / Yes
MISO [88]		Posterior mean estimates using Bayesian inference	Yes / Yes
RSEM [46]		Maximum likelihood estimation with EM algorithm	Yes / Yes

ALEXA-Seq stands for alternative expression analysis by sequencing; ERANGE, enhanced read analysis of gene expression; HTSeq, analyzing high-throughput sequencing data with Python; NEUMA, normalization by expected uniquely mappable area; IsoInfer, inference of transcripts from short sequence reads; rQuant, transcript quantification with RNA-seq data; IsoEM, transcript quantification by expectation maximization; MISO, mixture of transcripts; and RSEM, RNA-seq by expectation maximization.

1.5.3 Expression Normalization

The third step of the RNA-seq bioinformatics pipeline is expression normalization. Because of variations introduced in sequencing and bioinformatics processes, inter- or intra-sample comparisons of RNA-seq expression estimates can only be done after normalization. Most normalization methods for RNA-seq are based on scaling, in which the gene or transcript expression of any biological sample is normalized by multiplying or dividing by a fixed scaling factor. Therefore, the fundamental challenge for RNA-seq expression normalization is to estimate a set of robust scaling factors for samples in the dataset. **Table 5** lists commonly used RNA-seq normalization methods.

Several naïve methods such as RPM / FPM (reads / fragments per million mapped reads / fragments), median normalization, and upper-quartile normalization [96] are mathematically similar. RPM / FPM adjust expression estimates of each sample by the total number of mapped reads/fragments in the sample. Median and upper-quartile normalizations use the median and upper quartile read / fragment counts, respectively, of each sample as the substitute for the total mapped reads / fragments. With the Illumina sequencing protocol, longer genes or transcripts tend to produce a larger number of sequence fragments. Thus, some methods such as RPKM / FPKM (reads / fragments per kilobase per million mapped reads / fragments) [90] and TPM (transcripts per million) [46] further adjust expression estimates by gene or transcript length, which in turn enables both inter- and intra-sample comparisons. However, there exist limitations for the aforementioned normalization methods. For the RPKM / FPKM, gene or transcript length cannot be precisely defined. Also some methods such as RPM / FPM and RPKM / FPKM are sensitive to “extreme” datasets that have a small number of highly differentially

expressed genes. Therefore, a number of methods such as TMM (trimmed mean of M-values) [97] and RLE (relative log expression) [98] assume that most genes are not differentially expressed and use a robust estimate of library sizes as the scaling factors. Dillies *et al.* systematically evaluated a few normalization methods and recommended TMM or RLE as the most robust method for most RNA-seq data [99].

Table 5: Summary of Expression Normalization Methods for RNA-seq.

Normalization Method	Description
Median	Scaling by median of all counts
Quantile	Matching distributions of counts
RLE	Scaling by median ratio to median library
RPKM/FPKM	Scaling by library size and gene/transcript length
RPM/FPM	Scaling by library size
TMM	Scaling by estimate of relative RNA production
TPM	Scaling by mean length of expressed genes/transcripts
Upper Quartile	Scaling by upper quartile of all counts

RLE stands for relative log expression; RPKM/FPKM, reads/fragments per kilobase per million mapped reads/fragments; RPM/FPM, reads/fragments per million mapped reads/fragments; TMM, trimmed mean of M-values; and TPM, transcripts per million

1.6 Bioinformatics Challenges in the DIKW Hierarchy for NGS Data

The DIKW hierarchy is a framework that depicts the process from data to information, knowledge, and final wisdom (**Figure 3**). In the context of my thesis work, data refers to raw -omic data generated by NGS, information refers to extracted -omic features, knowledge refers to biomarkers or any patterns hidden in the -omic features, and wisdom refers to actionable knowledge for precision medicine. Linking components in this hierarchy requires various data analytics, and it is very challenging due to many

inherent properties in the -omic data such as data collection frequency, data quality, data dimensionality, data heterogeneity, and analytical pipeline complexity. These challenges motivate the specific aims of my thesis work, and more details of these challenges are discussed as follows:

Diverse Data Collection Frequency

For different -omic data modalities, data collection frequency varies tremendously. For example, a genome is invariant over a long period of time and often needs only one-time data acquisition, while other types of -omic data (e.g., transcriptomic and epigenomic) vary with environment, tissue types, and time that would require multi-time-point, multi-sample-source acquisition.

Inherent Data Quality Issues

In -omic data, quality issues are caused by a combination of biological, instrumental, and environmental factors such as sample contamination [100, 101], batch effects [102, 103], and low signal-to-noise ratios [104, 105]. These data quality issues may lead to wrong conclusion, but correcting these remains challenging.

High Dimensionality

An inherent challenge in data mining applications using -omic data is high data dimensionality. -Omic data often feature many dimensions or features (may be more than 10^4) much larger than the number of samples available [106-108]. This results in the "curse of dimensionality," a term which describes the phenomenon where the increasing dimensionality of the data exponentially increases the volume of the space needed to describe it, leading to increasingly sparse data filling the space [109].

Heterogeneous Data Type

In -omics, using underlying molecular fingerprints to characterize disease subtypes may require heterogeneous multi-omic data. For example, the integrative personal omics profile (iPOP) project has integrated multiple molecular expression profiles to uncover dynamic molecular changes between healthy and diseased states [110]. However, integrating multi-omic data is challenging because of variations in represented biological processes, technical and biological noise levels, identification accuracy, spatiotemporal resolution, and many other confounding factors [111].

High Analytical Pipeline Complexity

Ever since NGS became the most popular high-throughput assay for genomics, transcriptomics, and epigenomics, many bioinformatics solutions have been propose to extract -omic features from raw -omic data. Feature extraction for -omic data is a multi-step process and each step may have more than dozens of solutions. The high analytical complexity comes from all possible combinations of pipeline components for each NGS-based application. It has been known that significant variations in extracted -omic features exist that may be induced by data acquisition or pipeline variability. However, until now, no clear consensus about the choice of bioinformatics pipelines and its impact on downstream analysis has been established [36, 112, 113].

1.7 Structure of Dissertation

Motivated by the existing challenges mentioned in the previous section, I have formulated the overall objective and three specific aims of my thesis work as follows:

Overall Objective: To investigate and develop integrative bioinformatics approaches for extracting and discovering robust molecular knowledge for realizing future precision medicine.

Specific Aim #1 (Quality Control for Precision Medicine): To investigate and control the impact of bioinformatics pipelines on feature quality using RNA sequencing data.

Specific Aim #2 (Knowledge Discovery for Precision Medicine): To discover impactful biomarkers that facilitate disease subgroup assignment using NGS data.

Specific Aim #3 (Integrative Analysis for Precision Medicine): To integrate multiple sources of -omic data for improved disease subgroup assignment.

With the aim to promote the Precision Medicine Initiative, **Figure 4** summarizes the key areas where this dissertation improves the workflow from raw -omic data to disease subgroup assignment. The ultimate goal of precision medicine is to precisely classify patients into subgroups that share a common biological basis of diseases. The success of precision medicine relies on many building blocks, and my thesis work specifically addresses challenges from quality control, knowledge discovery, and integrative analysis perspectives. The quality control block ensures accurate -omic features being extracted from raw -omic data before discovering important knowledge for precision medicine (**Specific Aim #1**). The knowledge discovery block uses various models to identify key knowledge (e.g., distinguishing biomarkers) that facilitates precise disease subgroup assignment (**Specific Aim #2**). Finally, the integrative analysis block incorporates complimentary information from multiple sources that may lead to more precise disease subgroup assignment (**Specific Aim #3**).

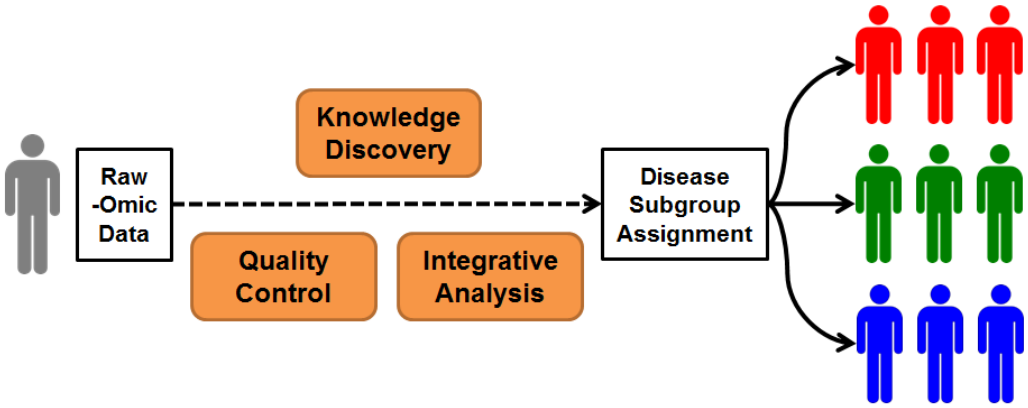


Figure 4: Overview of the Scope of This Dissertation. My thesis work aims to promote precision medicine by improving the workflow from raw -omic data generated by NGS to disease subgroup assignment. Many building blocks involve in this process, and my thesis work specifically focusses on quality control, knowledge discovery, and integrative analysis perspectives.

Chapter 2, Quality Control for Precision Medicine, addresses Specific Aim #1 by describing the experiment and evaluation-metric design that aims to establish RNA-seq expression analysis guidelines for accurate gene expression estimation. Two levels of investigation were conducted—pipeline component investigation and full pipeline investigation, and corresponding recommendations were provided. Chapter 3, Knowledge Discovery for Precision Medicine, addresses Specific Aim #2 by using statistical models and machine learning techniques to identify biomarkers that help classify patients into disease subgroups. Chapter 4, Integrative Analysis for Precision Medicine, addresses Specific Aim #3 by leveraging knowledge from multiple sources so as to improve the robustness of derived knowledge. Finally, Chapter 5, conclusion, provides concluding remarks, and highlights concrete deliverables that arose as a result of this dissertation. An outlook on future work for the Precision Medicine Initiative is also presented.

CHAPTER 2

QUALITY CONTROL FOR PRECISION MEDICINE

2.1 Introduction

The first objective of this dissertation was to investigate and control the impact of bioinformatics pipelines on feature quality using RNA-seq data. As introduced in Section 1.5, RNA-seq is a major branch of NGS that greatly facilitates transcriptomic research. **Figure 5** shows the schematic diagram from biological samples of interests to raw RNA-seq data generated by the NGS instrument. Among many RNA-seq applications, gene, transcript, or small RNA (e.g., miRNA) expression profiling and its downstream inferences have drawn much attention to researchers and clinicians and have brought huge impact to biology and medicine. As demonstrated in **Figure 6**, a typical RNA-seq expression analysis pipeline consists of sequence mapping, expression quantification, expression normalization, and one knowledge discovery step such as predictive modeling and differentially expressed gene detection. The first three components in the pipeline (i.e., sequence mapping, expression quantification, and expression normalization) are the focus of this chapter, and they are essential for extracting features (i.e., gene or transcript expression) from raw RNA-seq data. One key challenge of such the feature extraction process is that too many bioinformatics solutions are publicly available for each of the three components, and no consensus has been established about the impact of different choices of bioinformatics pipelines or pipeline components on downstream analysis and inferences. To ultimately promote precision medicine, it is important to control feature quality by investigating the performance of bioinformatics solutions for each pipeline

component or the pipeline as a whole. This is the central theme of this chapter and will be illustrated in detail throughout this chapter.

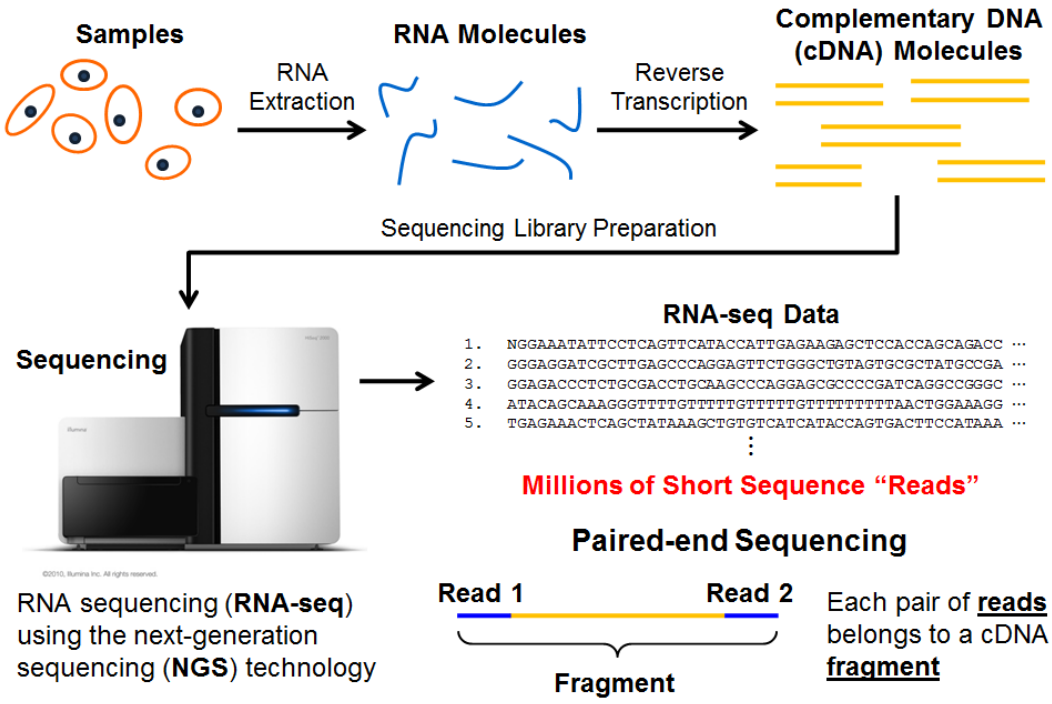


Figure 5: RNA Sequencing Workflow. The RNA sequencing, or RNA-seq for short, workflow starts from collecting samples of interests, followed by extracting RNA molecules in the samples, synthesizing cDNA molecules using reverse transcription techniques, preparing sequencing libraries, and finally running the NGS instrument to acquire raw RNA-seq data. The raw RNA-seq data are readings of RNA molecules in the samples. Currently, the most popular RNA-seq technique is called paired-end sequencing, in which both ends of the same fragment are sequenced.

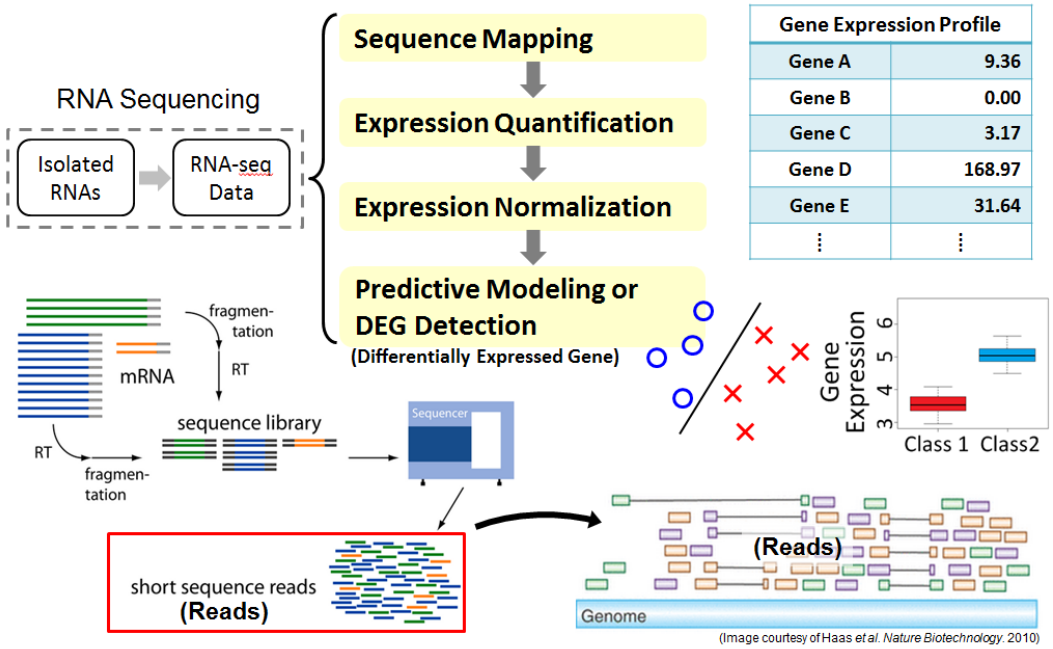


Figure 6: RNA-seq Expression Analysis Pipeline. The typical RNA-seq expression analysis pipeline includes sequencing mapping, which maps raw RNA-seq data, or reads, to the reference genome or transcriptome; expression quantification, which quantifies expression levels of each genes or transcripts; expression normalization, which normalized gene or transcript expression so that they become comparable with one another; and finally knowledge discovery, which identifies predictive, statistically significant biomarkers that may be crucial for the advancement of biology or medicine.

The rest of this chapter starts with the elaboration of feature extraction pipelines for RNA-seq data, followed by the introduction of evaluation metrics for pipeline performance assessment. I then use four case studies to showcase the investigation and control of the feature quality (i.e., the quality of gene or transcript expression) for precision medicine. The background, experimental design, datasets, and results of each case study are discussed based on their original publications [114-118]. A summary is provided at the end of this chapter with key accomplishments and innovations.

2.2 Feature Extraction Pipelines for RNA Sequencing Data

The thorough review of each pipeline component for RNA-seq expression analysis has been provided in Section 1.5. In this section, I mainly focus on introducing selected algorithms or methods I investigated in the four case studies as well as some complementary background to each component that has not been covered in Section 1.5.

2.2.1 Sequence Mapping

Inputs for Sequence Mapping Algorithms

Sequence mapping algorithms usually take three input files—sequence reads generated by the NGS instrument, reference sequences (e.g., a reference genome or transcriptome), and a genome annotation. The objective is to map sequence reads to reference sequences with the guidance of the genome annotation if desired.

The most well-known reference genome is UCSC (UC Santa Cruz) hg19, which mainly contains 24 primary chromosome contigs (i.e., chromosomes 1 to 22, X, and Y), 20 unplaced contigs (sequences with known chromosome but unknown chromosomal location), and 39 unlocalized contigs (sequences with unknown chromosome). Most of

my studies used UCSC hg19 as the reference genome. The reference transcriptome is typically extracted from the reference genome using genome annotation information. Several organizations or institutions have spent more than a decade working on annotating the human genome. Various annotating techniques have been developed and a variety of information sources have been utilized to provide the most informative and correct human genome annotation [119, 120]. In Case Study 1, my investigation included six well-known annotations, such as AceView Genes, Ensembl Genes, H-InvDB Genes, RefSeq Genes, UCSC Known Genes, and Vega Genes. The information sources and annotating strategies of each human genome annotation are summarized as follows:

AceView Genes [121]—The AceView annotation was downloaded from its website. The data sources of the AceView genes are mRNA sequences from GenBank and RefSeq as well as single pass cDNA sequences from dbEST and Trace. It summarizes all sequences into a comprehensive evidence-based gene annotation. It is a fully automatic process and uses heuristics to closely reproduce manual curation.

Ensembl Genes [122]—The Ensembl annotation was downloaded from its FTP site. The data sources of the Ensembl genes include (1) the automated Ensembl gene annotation pipeline "genebuild," (2) manually curated genes from the Havana Group at the WTSI, and (3) consensus coding sequences (CCDS). The final Ensembl genes result from clustering and merging these data sources.

H-InvDB Genes [123]—The H-InvDB annotation was downloaded from its website. H-InvDB genes are collected from six high-throughput sequencing projects [124]. It uses BLAST to map full-length cDNAs to the human genome, and then

annotates the genome based on clustering results. It assigns a standardized functional annotation to each H-InvDB transcript by manual curation.

RefSeq Genes [125]—The RefSeq annotation was downloaded from the UCSC Table Browser. The data sources of the RefSeq genes include all sequences submitted to the International Nucleotide Sequence Database Collaboration (INSDC), which consists of DDBJ, ENA, and GenBank. It combines an automatic genome annotation pipeline and a significant level of manual curation.

UCSC Known Genes [126]—The UCSC Known Genes annotation was downloaded from the UCSC Table Browser. The data sources of the UCSC known genes include protein data from Swiss-Prot / TrEMBL (UniProt) and the associated mRNA data from GenBank. It uses a fully automated process to annotate the genome.

Vega Genes [127]—The Vega annotation was downloaded from its website. The Vega database focuses on the browsing and maintenance of manually annotated data, including manually curated sequences from Havana, RIKEN, JGI, and Washington University.

Spliced Mapping versus Un-spliced Mapping

Spliced mapping refers to algorithms that split reads into segments in order to accommodate long gaps or introns in a read (e.g., TopHat and MapSplice); whereas un-spliced mapping refers to algorithms that align entire read sequences (e.g., Bowtie2, BWA, and Novoalign). For RNA-seq data, to directly map sequence reads to the reference genome, it is necessary to use spliced mapping algorithms because a read may map to exon-exon junctions that are equivalent to a long gap in the context of the reference genome. For Case Study 1, Impact of Genome Annotation Choice on Feature

Quality, I used two spliced mappers, TopHat and OSA, to map sequence reads to the human genome with the guidance of various genome annotations. A typical pipeline for spliced mapping is shown in **Figure 7**. TopHat is a spliced mapper that is widely used for mapping RNA-seq data to the reference genome [83]. Given a genome annotation, TopHat first maps short sequence reads to the reference transcriptome extracted from the reference genome, and then attempts to remap the unmapped reads from the previous stage to the reference genome. The mapping outputs from the two stages are then merged into the final output. If no genome annotation is given, TopHat uses a spliced technique (i.e., segmenting entire reads into smaller segments) to directly map sequence reads to the reference genome. OSA (Omicsoft sequence aligner) is a new spliced mapper that "improves mapping speed four- to ten-fold with better sensitivity and less false positives" compared to the TopHat, SoapSplice, and RUM [85]. It implements a similar mapping strategy as TopHat. Spliced mapping algorithms output mapping outcome in terms of genomic coordinates.

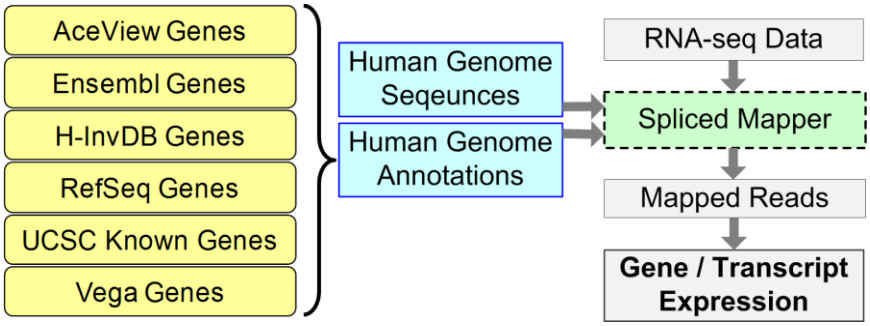


Figure 7: Typical RNA-seq Spliced Mapping Pipeline. A spliced mapper maps RNA-seq data to a reference genome with or without the guidance of a genome annotation. Different genome annotations define various sets of exon junction information that affect the spliced mapping output. The mapped reads can then be used to quantify gene or transcript expression.

Un-spliced mapping algorithms include both gapped alignment and un-gapped alignment. Thus, they can be applied to only transcriptomic mapping tasks. Un-spliced mapping algorithms output mapping outcome in terms of transcriptomic coordinates. Bowtie is one well-known un-spliced, un-gapped mapper.

In Case Study 4, Impact of Pipeline Choice on Feature Quality, our team extended the capability of un-spliced mapping by adding a coordinate translator that facilitates translating transcriptomic coordinates into genomic coordinates. With this module, any un-spliced mapper can be used for genomic mappings. **Figure 8** demonstrates both un-spliced and spliced RNA-seq mapping pipelines for Case Study 4.

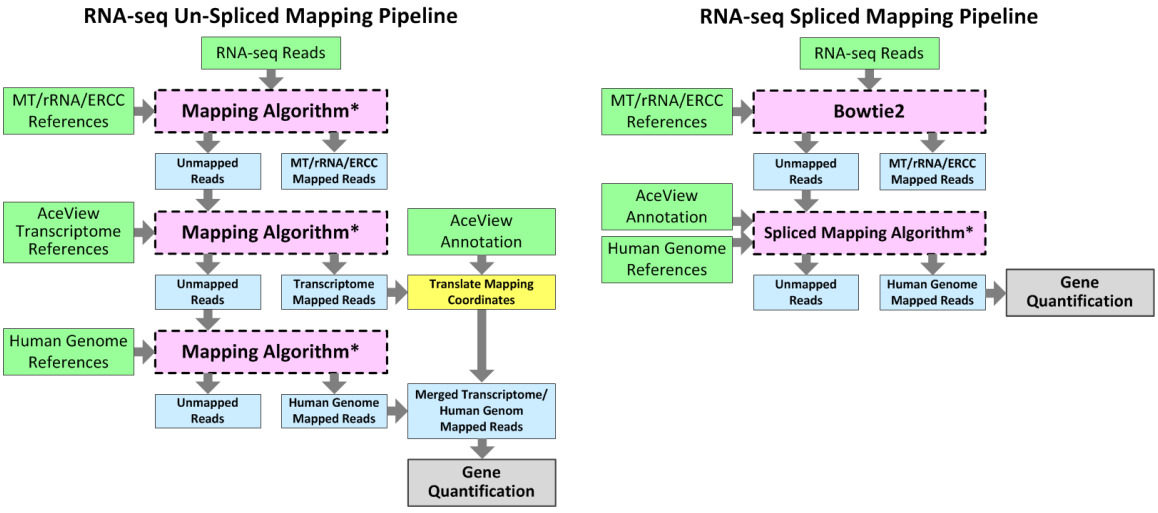


Figure 8: Un-spliced and Spliced RNA-seq Mapping Pipelines. Both pipelines map to the MT/rRNA/ERCC (mitochondria, ribosomal RNA, and External RNA Controls Consortium RNA spike-in mix) reference first. The un-spliced mapping pipelines combine transcriptome and genome mapping results into a single result for quantification (left). The spliced mapping pipelines either generate a single human genome mapping result or internally combine transcriptome and human genome mapping results into a single result for quantification (right). Thus, no additional merging step is required for spliced mapping pipelines.

Single-hit Reporting versus Multi-hit Reporting

Because the length of each sequence read is short (i.e., at the range of 36 bp to 300 bp for the current Illumina technology) and many repetitive regions exist in the human genome, one sequence read may map to multiple locations in the human genome. Depending on the algorithmic design and the application, a sequence mapper may report one or many alignments for this type of reads. Bowtie2, GSNAP, Novoalign, TopHat, and WHAM allow control over the number of reported mappings per read. By default, these algorithms typically report a single best mapping location per read. However, some quantification algorithms can use information about multiple ambiguous mapping locations to improve gene and transcript expression estimation. Thus, in addition to single-hit reporting, in Case Study 4, Impact of Pipeline Choice on Feature Quality, we also generated mapping results that reported up to 200 mapping locations per read.

Table 6 summarizes all mapping tools we applied in Case Study 4.

Table 6: RNA-seq Sequence Mapping Tools Studied in Case Study 4.

Sequence Mapping Tool	Mapping Strategy and Usage	Algorithmic Notes
Bowtie [73]	Un-spliced mapping to transcriptome or genome	Burrows-Wheeler transform and FM-index
Bowtie2 [75]		Burrows-Wheeler transform, FM-index-assisted seed alignment, dynamic programming
BWA [38]		Burrows-Wheeler transform
Novoalign		Commercial software, algorithm un-published
RUM [128]		Uses Bowtie in a multi-phase mapping to transcriptome and genome
WHAM [129]		Hash-based indexing and bitwise operations for fast alignment
GSNAP [82]	Spliced mapping to genome & un-spliced mapping to transcriptome or genome	Minimal sampling strategy, oligomer chaining for approximate alignment, sandwich dynamic programming
Magic [121, 130]		Strand-aware seed-and-extend alignment of read pairs, with at least 8 exact bases at the end of each aligned segment
MapSplice [84]	Spliced mapping to genome	Uses Bowtie for alignment, segmented mapping
OSA [85]		Two-stage transcriptome and genome alignment, segmented mapping
STAR [39]		Sequential maximum mappable seed search, and seed clustering and stitching
Subread [131]		Segmented seed-and-vote mapping
TopHat [83]		Uses Bowtie or Bowtie2 for alignment, segmented mapping

BWA stands for Burrows-Wheeler aligner; RUM, RNA-seq unified mapper; WHAM, Wisconsin's high-throughput alignment method; GSNAP, genomic short-read nucleotide alignment program; OSA, Omicsoft sequence aligner; and STAR, spliced transcripts alignment to a reference.

2.2.2 Expression Quantification

Count-based Quantification versus Model-based Quantification

Expression quantification algorithms can be roughly categorized into naïve count-based quantification and model-based quantification. For all case studies, I selected one or two representatives from each category and assessed their strengths and weaknesses through various experimental settings.

The most well-known count-based quantification tool is HTSeq. I used the htseq-count script from the HTSeq package to count the number of reads (or fragments in the paired-end sequencing case) that map to each gene as the gene expression estimate. For each mapped read or fragment, htseq-count determines the genes to which these reads or fragments associate. If a read or a fragment overlaps more than one gene, it provides three scenarios to resolve this ambiguous situation. For all case studies, I adopted the "intersection-nonempty" scenario if an ambiguity occurs [44].

Cufflinks is a popular model-based quantification tool that constructs graphical models describing how reads emit from each gene/transcript and estimates gene/transcript expression using the maximum likelihood estimation technique. It is capable of both assembling transcripts and quantifying gene or transcript expressions. In my case studies, I disabled the assembly function and provided the genome annotation GTF file that contains description of targeted genes and their structures as a quantification reference. I used Cufflinks with default setting except enabling sequencing bias correction and multi-mapped reads correction [47]. Information from multi-hit reads is important for model-based quantification tools. These algorithms use multi-hit read information to estimate gene or transcript expression more accurately.

Transcriptomic Mapping versus Genomic Mapping

As briefly touched upon in the previous section, RNA-seq data can be mapped to either the reference genome or the reference transcriptome. Therefore, some quantification algorithms were specifically designed for quantifying genomic mapping outputs, while others were designed for quantifying transcriptomic mapping outputs.

For Case Study 2, Impact of Expression Quantification Choice on Feature Quality, I applied Cufflinks, HTSeq, and MISO to quantify gene and transcript expression for the sequence alignment reported in genomic coordinates. The Cuffdiff 2 program in the Cufflinks package produces gene and transcript expression estimates in terms of read count and FPKM [90] (i.e., fragments per kilobase of exon per million fragments mapped) estimates. The htseq-count program in the HTSeq package generates read count estimates only for genes, while MISO provides read count estimates for both genes and transcripts. For sequence alignment reported in transcriptomic coordinates, I used RSEM, eXpress, and MMSEQ to quantify gene and transcript expression. These three quantification algorithms are able to quantify both genes and transcripts and produce both read count and FPKM estimates. In addition, RSEM provides in-house TPM (i.e., transcripts per million) estimates. For read count estimates, I applied the same trimmed mean of M-values normalization method (TMM) [97] to eliminate the effect of the normalization factor when computing evaluation metrics.

Quantification Pipeline Compatibility

Mapping results from alignment pipelines were not always compatible with quantification tools. Cufflinks requires alignment files to be sorted by alignment coordinates and multi-hit reads to be annotated with the ‘NH’ tag in the attribute field of

the SAM (sequence alignment/map format) file. HTSeq requires that the alignment files are sorted by read name and that the 'NH' tag is not present in the SAM file. RSEM only quantifies transcriptome mapping, i.e., reads mapped and reported in transcriptomic coordinates. Moreover, RSEM only handles un-gapped alignments. Thus, filtering is required to remove gapped alignments. Because of these requirements, pre-processing alignment outputs before quantification is needed. For Case Study 4, Impact of Pipeline Choice on Feature Quality, in total, twenty alignment pipelines, including spliced, un-spliced, single-hit, and multi-hit pipelines, were suitable for count-based quantification. Sixteen alignment pipelines were suitable for Cufflinks, and only ten were suitable for RSEM. RSEM is specifically designed to work well with the Bowtie alignment tool. Thus, we included this embedded mapping and quantification pipeline in RSEM.

The key characteristics of these quantifiers are summarized in **Table 7**.

Table 7: RNA-seq Expression Quantification Tools Studied in Case Study 4.

Expression Quantification Tool	Mathematical Model & Estimation Method	Quantification Targets
HTSeq [44]	Count-based; accumulated counts	Gene
Magic quantifier (for Magic mapper only) [121, 130]		Gene, Transcript
Subread featureCounts (for Subread mapper only) [132]		Gene, Transcript
RUM quantifier (for RUM mapper only) [128]		Gene, Transcript
Cufflinks [33]	Poisson model; maximum likelihood with maximum a posteriori estimates using Bayesian inference	Gene, Transcript
RSEM [46]	Poisson model; maximum likelihood estimation using expectation-maximization algorithm	Gene, Transcript

2.2.3 Expression Normalization

RNA-seq expression normalization enables inter- or intra-sample comparison. Generally, normalization methods correct the library size (i.e., the total number of mapped reads in a sample), which is the primary source of inter-sample variability. We used seven normalization methods (**Table 8**): FPM (fragments per million mapped reads), FPKM (fragments per kilobase of gene length per million mapped reads), median, upper quartile, RLE (relative log expression), TMM (trimmed mean of M-values), and expression index (specific to the Magic pipeline). I describe each of these normalization methods in the context of Case Study 4, but they are all applicable to any other experimental settings.

Table 8: RNA-seq Expression Normalization Methods.

Normalization Method	Description
Reads/Fragments Per Million mapped reads/fragments (RPM/FPM) [99]	Scaling by library size
Reads/Fragments Per Kilobase per Million mapped reads/fragments (RPKM/FPKM) [90]	Scaling by library size and gene/transcript length
Median [99]	Scaling by median of all counts
Upper Quartile (UQ) [99]	Scaling by upper quartile of all counts
Relative Log Expression (RLE) [98]	Scaling by median ratio to median library
Trimmed Mean of M-values (TMM) [97]	Scaling by estimate of relative RNA production
Expression Index (Eindex) [121, 130]	Magic pipeline only, scaling and thresholding to identify low-expression genes

In Case Study 4, Impact of Pipeline Choice on Feature Quality, the raw count of a sample is defined as $x_{s,n,k}$ where $s \in \{A, B, C, D\}$ indicates the sample, $n = 1 \ldots N$ indicates the replicate, and $k = 1 \ldots K$ indicates the gene. For the benchmark dataset in

Case Study 4 (will be introduced in detail in Section 2.4.4), N = 4 and K = 55,874. Since genes with zero counts contribute negatively to normalization performance, we first identified and used only non-zero genes during normalization. Given that the mean of the counts for gene k, and sample s over all replicates is

$$\bar{x}_{s,\cdot,k} = \frac{1}{N}\sum_{n=1}^{N} x_{s,n,k} \,, \tag{1}$$

we defined the set of "present" genes to be

$$K_p \in \left\{k \middle| \left(\bar{x}_{A,\cdot,k} > 1 \vee \bar{x}_{B,\cdot,k} > 1\right) \wedge \bar{x}_{C,\cdot,k} > 1 \wedge \bar{x}_{D,\cdot,k} > 1 \right\}. \tag{2}$$

The total count of present genes for a given sample s and replicate n is

$$x_{s,n} = \sum_{k \in K_p} x_{s,n,k} \,, \tag{3}$$

and the average total count of present genes for all data from a single site is

$$\bar{x} = \frac{1}{4}\frac{1}{N}\sum_{s}\sum_{n=1}^{N} x_{s,n} \,. \tag{4}$$

Thus, we can define the FPM normalization as

$$y_{s,n,k}^{FPM} = \frac{x_{s,n,k}\cdot\bar{x}}{x_{s,n}} \,. \tag{5}$$

Similarly, if we define $\tilde{x}_{s,n}$ and $\hat{x}_{s,n}$ as the median and upper quartile of counts, respectively, of all present genes in sample s and replicate n, then

$$\tilde{x} = \frac{1}{4}\frac{1}{N}\sum_{s}\sum_{n=1}^{N} \tilde{x}_{s,n} \quad \text{and} \quad \hat{x} = \frac{1}{4}\frac{1}{N}\sum_{s}\sum_{n=1}^{N} \hat{x}_{s,n} \,. \tag{6}$$

Median and upper quartile normalizations are then defined as

$$y_{s,n,k}^{Med} = \frac{x_{s,n,k}\cdot\tilde{x}}{\tilde{x}_{s,n}} \quad \text{and} \quad y_{s,n,k}^{UQ} = \frac{x_{s,n,k}\cdot\hat{x}}{\hat{x}_{s,n}} \,. \tag{7}$$

For FPKM normalization, we defined the length of a gene k as ℓ_k, which is the length of the union of all exons related to the gene as defined by the AceView transcriptome. The original formulation of FPKM arbitrarily used scaling factors of $1\text{x}10^3$ for the gene length and $1\text{x}10^6$ for total number of mapped reads. In order to maintain

comparable dynamic range among all normalization methods, we instead scaled by the average gene length and average total count for all present genes. The average length of all present genes is

$$\bar{\ell} = \frac{1}{|K_p|}\sum_{k\in K_p} \ell_k \ . \quad \textbf{(8)}$$

Thus, the rescaled FPKM normalization is

$$y_{s,n,k}^{FPKM} = \frac{x_{s,n,k}\cdot\bar{\ell}\cdot\bar{x}}{x_{s,n}\cdot\ell_k} \ . \quad \textbf{(9)}$$

TMM and RLE normalizations are similar to the FPM normalization, but introduce an extra scaling factor. We used the edgeR package in R to estimate scaling factors for each sample and replicate [65, 97]. The TMM method selects a reference sequence library from the pool of samples and then calculates gene-wise log expression ratios (M-values) and gene-wise average log expression values (A-values) between the target library and the reference library. Extreme numbers in M-values and A-values are trimmed and the scaling factor for the target library is the weighted average of remaining M-values. The RLE method determines a scaling factor by first defining median library size as the gene-wise geometric mean across samples and replicates [98]. The median ratio of each sequence library to the median library is taken as the scaling factor. TMM and RLE normalizations are then defined as:

$$y_{s,n,k}^{TMM} = \frac{x_{s,n,k}\cdot\bar{x}}{x_{s,n}\cdot\hat{f}_{s,n}^{\ TMM}} \quad \text{and} \quad y_{s,n,k}^{RLE} = \frac{x_{s,n,k}\cdot\bar{x}}{x_{s,n}\cdot\hat{f}_{s,n}^{\ RLE}} \ , \quad \textbf{(10)}$$

where $\hat{f}_{s,n}^{\ TMM}$ and $\hat{f}_{s,n}^{\ RLE}$ are the scaling factors.

2.3 Evaluation of Feature Extraction Pipeline Performance

To evaluate the performance of feature extraction pipelines in terms of gene or transcript expression quality, I designed various evaluation metrics for both sequence mapping and expression quantification / normalization results.

2.3.1 Evaluation Metrics for Sequence Mapping

I defined two evaluation metrics for sequence mapping outputs. The first metric is based on the categorization of read mapping outcomes. For paired-end sequencing, the categories include uniquely paired reads, uniquely mapped singletons, non-uniquely paired reads, non-uniquely mapped singletons, and unmapped reads. For single-end sequencing, the categories are simpler, including only uniquely mapped reads, non-uniquely mapped reads, and unmapped reads. The second metric is the percentage of the number of reads that map to the annotated and un-annotated genomic sequences.

2.3.2 Evaluation Metrics for Expression Quantification and Normalization

I designed and defined various evaluation metrics for assessing the quality of gene or transcript expression estimates. Most metrics rely on normalized expression, while a few are independent of normalization methods. The evaluation metrics include (1) accuracy, which measures the gene- or transcript-level deviation between expression estimates based on RNA-seq and those based on the predefined ground truth (e.g., simulation and qPCR); (2) precision, which measures the gene- or transcript-level dispersion of expression estimates across replicate libraries; (3) reproducibility, which measures sample-level variation under similar experimental conditions; and (4) reliability, which measures gene- or transcript-level intra-sample consistency among replicate libraries.

Accuracy Measured as Deviation from the Ground Truth

As illustrated in **Figure 9**, accuracy describes how close measurements to the reference value (or true value). In my case studies, measurements are gene or transcript expression extracted from raw RNA-seq data and the reference value can be either simulated expression or qPCR-based expression (with the assumption that qPCR is the ground truth, which is commonly acknowledged in the bioinformatics community). For this metric, smaller deviation indicates higher accuracy. The rest of this section will elaborate more on several variants of this metric for different case studies.

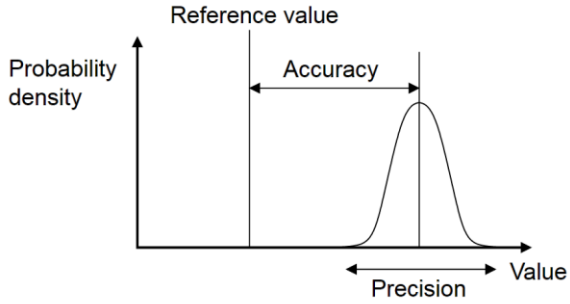


Figure 9: Accuracy and Precision of the Measurement System. The accuracy of a measurement system describes the closeness of measurements to the true value, and the precision of the measurement system captures the repeatability of consecutive measurements. The bell curve shows the distribution of many measurements for a single quantity. The accuracy depends on the distance between mean measurements of the quantity and the reference value, while the precision depends on the dispersion of these measurements.

In Case Study 1, Impact of Genome Annotation Choice on Feature Quality, I used the mean absolute deviation (MAD) and the root-mean-squared error (RMSE) to quantify the deviation between RNA-seq-based log ratios and qPCR-based log ratios.

$$MAD = \frac{\sum_n |log2(FC_{RNA-seq}) - log2(FC_{qPCR})|}{n}; \quad \textbf{(11)}$$

$$RMSE = \sqrt{\frac{\sum n\left[log2(FC_{RNA-seq}) - log2(FC_{qPCR})\right]^2}{n}}, \quad \textbf{(12)}$$

where *FC* stands for fold change and *n* is the number of genes or transcripts in the pool.

In Case Study 2, Impact of Expression Quantification Choice on Feature Quality, I quantified deviation by computing the normalized RMSE between estimated counts from quantification pipelines and true counts from simulation. Since RMSE is not scale-invariant, to adjust the sequencing depth effect, I normalized the RMSE by dividing the original RMSE by 10 for 100-million-read cases, by 5 for 50-million-read cases, and by 1 for 10-million-read cases.

In Case Study 4, Impact of Pipeline Choice on Feature Quality, we computed inter-sample log ratios for each gene for both RNA-seq pipeline-produced gene expression and qPCR-produced gene expression, and the deviation was defined as the difference between inter-sample log ratios from both sources. The formulation of this process is as follows:

We defined RNA-seq assayed gene expression as $x_{s,n,k}$ and qPCR assayed gene expression as $y_{s,n,k}$ where $s \in \{A, B, C, D\}$ indicates the sample, $n = 1 \ldots N$ indicates the replicate, and $k = 1 \ldots K$ indicates the gene (for the PrimePCR set, N = 1; K = 10,222). The mean expression of a qPCR gene for gene *k*, and sample *s* over all replicates is

$$\bar{y}_{s,\cdot,k} = \frac{1}{N}\sum_{n=1}^{N} y_{s,n,k} \ . \quad \textbf{(13)}$$

Given samples A and B, the absolute log-ratio deviation between RNA-seq-based expression and qPCR-based expression is

$$\Delta_{\frac{A}{B},k} = \left| log_2\left(\frac{\bar{x}_{A,\cdot,k}}{\bar{x}_{B,\cdot,k}}\right) - log_2\left(\frac{\bar{y}_{A,\cdot,k}}{\bar{y}_{B,\cdot,k}}\right)\right|, \quad \textbf{(14)}$$

and the final deviation was defined as the median of all $\Delta_{\frac{A}{B},k}$, $k = 1 \ldots K$.

Precision Measured as Variation across Replicate Libraries

As also illustrated in **Figure 9**, precision describes the dispersion of multiple measurements for the same quantity. Since gene expression differs by orders of magnitude, statistics that capture normalized dispersion are needed. One popular statistic that fits this objective is the coefficient of variation (CoV), which is defined as the ratio between the standard deviation and the mean of replicate measurements. For this metric, smaller CoV indicates higher precision. Note that this metric will only work when replicate libraries are available. The rest of this section will elaborate more on several variants of this metric for different case studies.

In Case Study 1, Impact of Genome Annotation Choice on Feature Quality, I removed genes that are absent (i.e., have zero expression) in all replicate libraries, calculated CoV for each remaining gene, and finally computed average CoV across all targeted genes or transcripts as follows:

$$Average\ CoV = \frac{\sum_{i=1}^{n} S_i/\bar{x_i}\,|_{condition\ 1} + \sum_{i=1}^{n} S_i/\bar{x_i}\,|_{condition\ 2}}{2n} \quad \textbf{(15)}$$

where S_i and $\bar{x_i}$ are the sample standard deviation and mean of expression estimates across replicate libraries with the same biological condition, respectively, and *n* is the number of targeted genes or transcripts. I applied the same metric to different sets genes and transcripts in Case Study 1.

In Case Study 2, Impact of Expression Quantification Choice on Feature Quality, I computed the condition-wise CoV for each gene and then summarized these CoVs into the gene-wise CoV. The same technique applies to the transcript-wise CoV analysis.

In Case Study 4, Impact of Pipeline Choice on Feature Quality, we computed the CoV for each gene and each sample across four replicate libraries as follows:

$$CoV_{s,k} = \frac{sd(x_{s,\cdot,k})}{\bar{x}_{s,\cdot,k}}, \quad \textbf{(16)}$$

where $k = 1 \ldots K$ indicates the gene, and $s \in \{A, B, C, D\}$ indicates the sample. We only consider K = 10,222 genes that were expressed as non-zero in all conditions. We then computed the median of all CoV derived from all genes and all samples as the final measure of precision.

Metric based on Concepts of Accuracy and Precision

In Case Study 2, Impact of Expression Quantification Choice on Feature Quality, I calculated the gene-wise CoV across all replicate libraries for the simulated dataset that serves as the ground-truth measurement for replicate variation. I then examined the percent error when comparing the gene-wise CoV from various quantification pipelines with that from the simulated dataset. Finally, I counted the cumulative number of genes or transcripts with percent errors of the CoV from 0% to 5%, from 0% to 10%, from 0% to 15%, and so on, up to from 0% to 100%. The same technique applied to the transcript-wise CoV analysis.

Reproducibility Measured as Inter-sample Correlation

According to the definition described in [133], reproducibility refers to "the variation in measurements made on a subject under changing conditions." The changing conditions may involve with different instruments for measurements, different measurement time points, different observers, and many other factors. In the context of my case studies, since no two replicate libraries of RNA-seq are exactly identical to each other due to various biological, chemical, instrumental, and experimental factors, the assessment of variation between replicate libraries falls within the scope of

reproducibility. I chose to use the inter-sample Spearman correlation coefficient as the measure of reproducibility, which is a commonly applied metric for reproducibility. Higher Spearman correlation coefficients indicate higher reproducibility.

This metric was applied to only Case Study 4, Impact of Pipeline Choice on Feature Quality. In Case Study 4, we computed the pairwise Spearman correlation coefficient between replicate libraries for the same sample s, $s \in \{A, B, C, D\}$. Since each sample has four replicate libraries, there are six pairwise comparisons for each sample and 24 comparisons in total for the entire dataset. We then computed the median of all 24 Spearman correlation coefficients as the final measure of reproducibility.

Reliability Measured as Intraclass Correlation

The reliability of a measurement system can be assessed by the intraclass correlation coefficient (ICC) [133, 134]. Conceptually, ICC is applicable to measurements that can be organized into groups, and it describes how similar measurements of the same group are to one another. Modern ICC definition borrows the framework of analysis of variance (ANOVA), or more specifically ANOVA with random effects [134]. The type of ANOVA depends on the experimental design and generally follows the definition in [134]. *ICC*(1,1) and *ICC*(1,k) are based on the one-way random effects model and are applicable to the case that each group is assessed by a different set of *k* raters randomly selected from a larger population of raters. *ICC*(2,1) and *ICC*(2,k) are based on the two-way random effects model and are applicable to the case that a random sample of *k* raters is preselected from a larger population and each rater assesses each group exactly once (i.e., each rater assesses *n* groups altogether). *ICC*(3,1) and *ICC*(3,k) are based on the two-way mixed effects model and are applicable to the case

that each group is assessed by each of the same *k* raters, who are the only raters in the population. The second parameter in *ICC*([1,2,3],[1,k]) denotes whether the ICC is to measure the reliability of a single measurement or the average of k measurements.

For my case studies with replicate libraries for each sample, *ICC*(1,1) or *ICC*(1,k) fitted my objective since for a specific gene *g*, gene expression of replicate libraries for different samples (or different groups in the previous context) were not assessed under exactly the same conditions (or assessed by the same raters in the previous context). Finally, *ICC*(1,k) was my final choice since replicate libraries are available for most experiments. Mathematically, a one-way random effects model can be formulated as

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij} , \quad \textbf{(17)}$$

where Y_{ij} is the i^{th} observation in the j^{th} group, μ is an unobserved overall mean, α_j is the group-specific random effect with zero mean and variance ${\sigma_\alpha}^2$, and ϵ_{ij} is an unobserved noise term with zero mean and variance ${\sigma_\epsilon}^2$. The formula for *ICC*(1,k) is as follows:

$$ICC(1,k) = \frac{BMS - WMS}{BMS}, \quad \textbf{(18)}$$

where *BMS* stands for the between groups mean square and has the expected mean square of $k\sigma_\alpha^2 + \sigma_\epsilon^2$, and *WMS* stands for the within group mean square and has the expected mean square of σ_ϵ^2. Higher ICC indicates higher reliability.

This metric was applied to only Case Study 4, Impact of Pipeline Choice on Feature Quality. In Case Study 4, we calculated ICC for each gene, and then computed the median of all ICCs as the final measure of reliability.

2.4 Case Study

To demonstrate my approach for quality control for precision medicine, I detail four case studies in this section, including the impact of genome annotation, expression

quantification, normalization, and the overall pipeline choice on feature (i.e., gene or transcript expression) quality. The background, experimental design, datasets, and results of each case study are discussed based on their original publications [114-118].

2.4.1 Impact of Genome Annotation Choice on Feature Quality

2.4.1.1 Background

RNA-seq is a major branch of the NGS technology that studies the transcriptome [135]. One aspect of transcriptomic research is quantification of expression levels for various genomic elements such as genes and transcripts [90]. Acquiring a transcriptomic expression profile requires knowledge about the location of genomic elements in the context of the reference genome, and such the knowledge is provided by a process called genome annotation. Multiple human genome annotations are publicly available, including, but not limited to, the AceView database [121] and the RefSeq database [125]. Thus, it is necessary to study the impact of genome annotation choice on gene or transcript expression quality derived from RNA-seq data.

Genome annotation is a dynamic process that defines coordinates for each genomic element with respect to the genome sequence. Such a process bridges the gap between DNA or RNA sequences and biological functions [136]. Integration of a genome annotation with mapping information from RNA-seq short sequence reads enables quantification of genomic elements. Each genome annotation project adopts different annotation strategies and information sources. Thus, high variation exists among publicly available annotations in terms of the comprehensiveness of annotated genomic elements. Some annotation strategies rely on computer-based prediction, resulting in more complex gene models that contain more putative genomic elements. Other annotation strategies

rely on evidence-based methods, that is, methods that require more manual curation, leading to simpler gene models with a fewer number of genes and transcripts.

I compared six human genome annotations from various databases, including the AceView database [121], the Ensembl database [122], the H-InvDB database [123], the RefSeq database [125], the UCSC Known Genes database [126], and the Vega database [127]. The key characteristics of each genome annotation are summarized in **Table 9**, in which annotations are ordered by decreasing complexity from left to right. The term "complexity" describes the primary differentiating characteristic among the genome annotations. I defined the complexity of a human genome annotation to be proportional to the number of genes, transcripts, and exons. This definition enabled me to investigate the relationship between the measure of genome annotation complexity and the outcome of the RNA-seq expression analysis pipeline. I hypothesized that a more complex genome annotation is more difficult for RNA-seq mapping and quantification because of the difficulties of determining a best possible mapping from multiple candidate mappings and assigning unresolved ambiguous mappings to their correct genomic elements.

Any relationship between genome annotation complexity and gene expression quality could be informative in guiding the selection of a genome annotation for various gene / transcript expression-based studies using RNA-seq data. Currently no guidelines for selecting a genome annotation for RNA-seq expression analysis are available, and the effect of genome annotation choice on downstream data analysis is still unclear. The focus of this study was to obtain some insights into the impact of human genome annotation choice on RNA-seq expression estimates.

Table 9: Properties of Various Human Genome Annotations.

	Genome Annotation					
	AceView Genes	*H-InvDB Genes*	*Ensembl Genes*	*Vega Genes*	*UCSC Known Genes*	*RefSeq Genes*
Version	2010	8.0	67	48	-	-
Database Downloaded Date	Nov. 12, 2011	Apr. 20, 2012	May 1, 2012	June 26, 2012	Dec. 21, 2011	July 23, 2012
# of Genes	72,376	43,893	48,817	44,880	28,423	23,731
# of Transcripts	259,426	236,861	177,858	158,835	75,725	41,099
# of Exons	678,503	542,099	534,400	493,509	273,711	227,710
Average # of Transcripts per Gene	3.58	5.40	3.64	3.54	2.66	1.73
Maximum # of Transcripts per Gene	119	885	82	77	129	77
Annotated Percentage (%)						
Gene	52.93	45.09	49.61	48.29	44.28	40.17
Exon	5.70	3.72	3.63	3.53	2.70	2.27
Coding Sequence	1.71	1.43	1.14	1.05	1.13	1.07

The annotated percentage is the total length of all genomic elements (gene, exon, or coding sequence) over the entire length of the human genome.

2.4.1.2 Experimental Design

This case study aims to provide insights into the effect of different choices of the human genome annotation on the variation in RNA-seq expression estimates. I proposed a complexity measure (referring to the previous section) that relates observations in downstream RNA-seq analysis to the trend of genome annotation characteristics. The typical pipeline for RNA-seq expression analysis includes sequence mapping, expression quantification, expression normalization, and calling DEGs. As shown in **Figure 10**, in this case study, I used two publicly available RNA-seq datasets that provide a list of DEGs and qPCR validation information. I mapped short sequence reads to the human reference genome with two spliced mappers, OSA [85] and TopHat [83]. Alignment

outputs of both tools were quantified by htseq-count [44] to acquire gene expression estimates in terms of the read counts. Since OSA has embedded quantification and TPM normalization [46] in its package, I used Cufflinks [47] to quantify TopHat alignment outputs only and then obtained gene / transcript expression in terms of FPKM-normalized values [90]. Given the read counts data from htseq-count, I applied the edgeR package in R [65] to call DEGs between treatment and control samples. For TPM or FPKM expression estimates, I calculated fold changes between treatment and control samples and then compared these fold changes to external qPCR validation results provided by the original studies. I proposed several evaluation metrics for each analytical step to demonstrate performance variation induced by the genome annotation complexity.

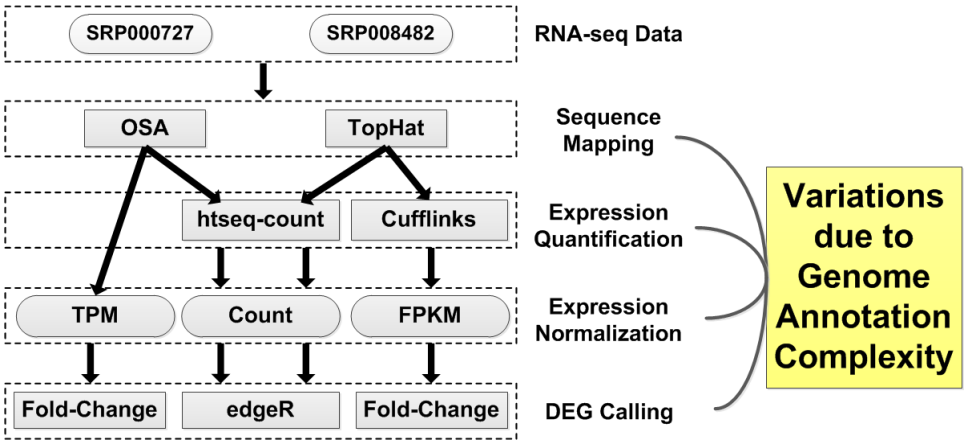


Figure 10: Workflow for Chapter 2, Case Study 1. The five dashed boxes correspond to five steps in the RNA-seq data analysis pipeline. I applied two sequence mapping tools and two expression quantification tools to estimate gene / transcript expression with normalization methods of count, TPM, or FPKM. The fold-change method and the edgeR tool were used to infer DEGs. At each analytical step, I assessed variations resulting from the choice of human genome annotation.

2.4.1.3 Datasets

I downloaded two publicly available RNA-seq datasets from the NCBI Sequence Read Archive (SRA) repository. The first dataset (accession number: SRP008482) investigates how thrombin treatment affects endothelial function in terms of gene expression profiles. In general, thrombin can stimulate endothelial cells and regulate the expression, release and activation of a number of biological mediators [137]. The targeted biological samples are "human pulmonary microvascular endothelial cells (HMVEC-L)" with two conditions—control (two technical replicates) and thrombin treatment for six hours (three technical replicates). The sequencing platform was Illumina HiScanSQ with the sequencing depth at about 50 million read pairs for each technical replicate and the read length of 101 bp. The study also validated expression fold changes of three genes (CELF1, FANCD2, and TRAF1) between treated and control samples using a qPCR assay. Such qPCR information is considered the ground truth and is useful for validating and evaluating RNA-seq expression estimates.

The second dataset (accession number: SRP000727) studies alternative transcript regulation in human tissue transcriptomes [138]. It profiled 16 tissue transcriptomes, and two MAQC (microarray quality control) samples were included in the study. The two MAQC samples are Ambion Human Brain Reference RNA (HBRR) and Stratagene Universal Human Reference RNA (UHRR). The study includes four technical replicates for the HBRR sample and 3 technical replicates for the UHRR sample. This is an older sequencing dataset which used Illumina Genome Analyzer to generate single-end reads with the read length of 36 bp. Each technical replicate has only 2.5 million reads. The merit of this dataset is that the qPCR results are publicly available through the MAQC

project. Fold changes of 1,044 genes from the TaqMan qPCR assay again provide an external ground truth for evaluating RNA-seq expression estimates.

2.4.1.4 Results and Discussion

Complexity of human genome annotations

Table 9 summarizes several important statistics for each genome annotation. I ranked the genome annotation based on the number of genes, transcripts, and exons. Ranking the set of human genome annotations (i.e., AceView, H-InvDB, Ensembl, Vega, UCSC, and RefSeq) by decreasing number of genes resulted in ranks of (1, 4, 2, 3, 5, and 6). In other words, AceView was ranked at 1 because it had the most number of genes, while H-InvDB was ranked at 4. Similarly, ranking the set of human genome annotations by decreasing number of transcripts and exons resulted in identical ranks of (1, 2, 3, 4, 5, and 6) and (1, 2, 3, 4, 5, and 6), respectively. I then defined the complexity rank of the genome annotation to be proportional to the average of these three ranks. The average ranks of these genome annotations are (1, 2.67, 2.67, 3.67, 5, and 6). I used the mode of ranks to break ties (e.g., the H-InvDB and Ensembl annotations both ranked 2.67). Thus, the human genome annotations were ordered by decreasing complexity as AceView, H-InvDB, Ensembl, Vega, UCSC, and RefSeq. The annotated percentage of each genome annotation generally follows the trend of complexity as demonstrated in **Figure 11**. For the average number of transcripts per gene and the maximum number of transcripts per gene, annotations generally have the same trend as the complexity measure. However, the H-InvDB annotation deviates from this trend, containing on average 50% more transcripts per gene compared with the most complex AceView annotation.

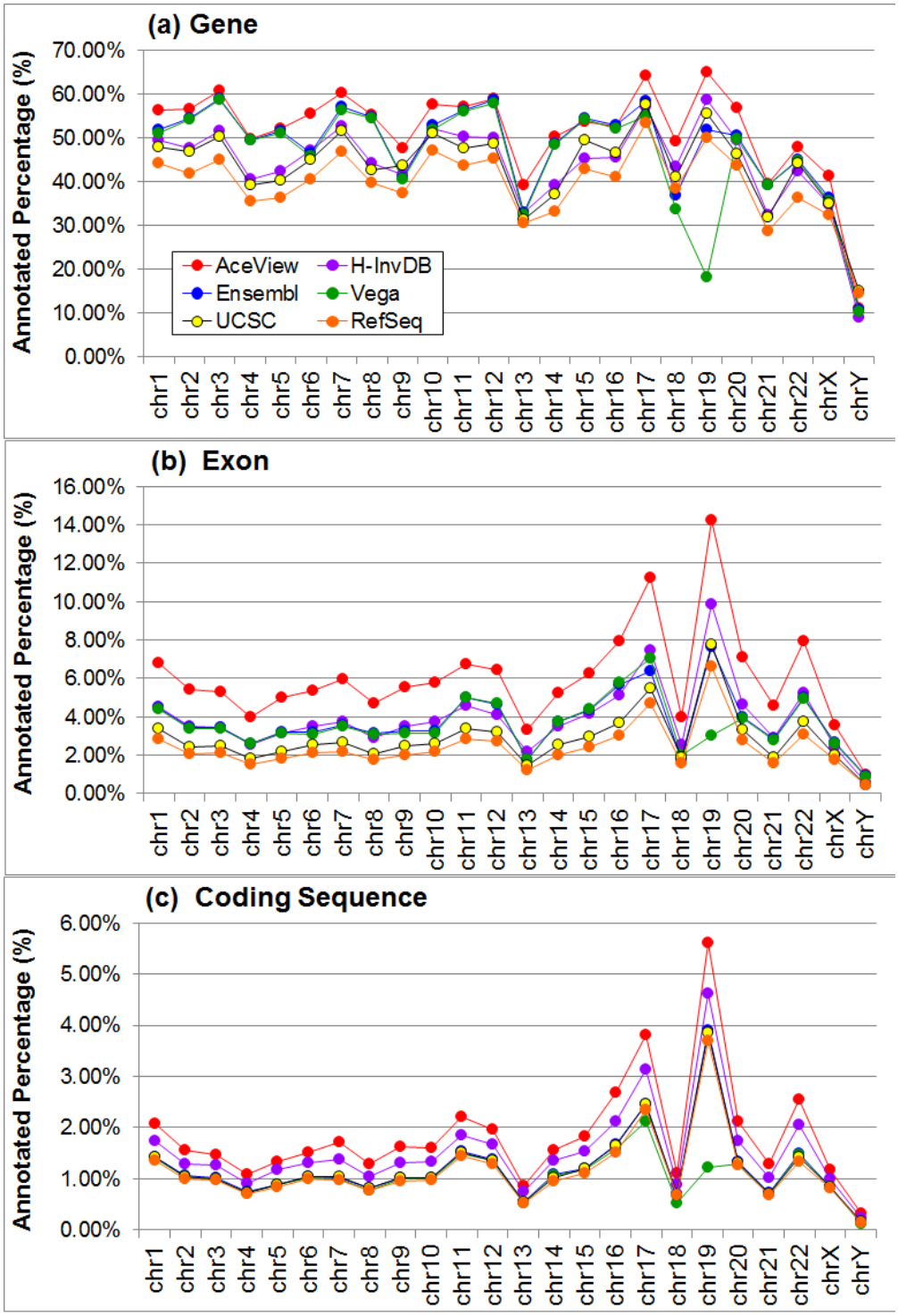


Figure 11: Annotated Percentage per Chromosome. For each genome annotation, the annotated percentage of each chromosome is demonstrated on (a) the gene level, (b) the exon level, and (c) the coding sequence level. The AceView annotation usually has the highest annotated percentage for all chromosomes and all levels of comparison.

Effect of human genome annotation complexity on mapping

I proposed two metrics to assess the effect of genome annotation complexity on sequence mapping. I first examined read mapping information and classified them into three categories for single-end sequencing samples or into five categories for paired-end sequencing samples. I used OSA alignment outputs as an example to demonstrate the impact of genome annotation choice on read mapping. For both the shorter read length single-end sequencing samples (1×36 bp; SRP000727) and the longer read length paired-end sequencing samples (2×100 bp; SRP008482), I observed similar results (**Figure 12**). The RefSeq annotation consistently had the highest percentage of uniquely mapped reads and uniquely paired reads in the single-end case and paired-end case, respectively. Note that the percentage of unmapped reads was similar for all annotations. The percentage of non-uniquely mapped reads or read pairs increased as the genome annotation becomes more complex. Outlying cases existed (e.g., the Vega annotation had the lowest percentage of uniquely paired reads in paired-end sequencing samples), but the observed trend still followed the complexity measure. From **Table 9**, more complex annotations generally annotate more genes and transcripts, and thus, they increase the possibility of ambiguous mappings. These ambiguous mappings are more difficult to resolve for identifying the best mapping, which directly translates to the increase in the percentage of non-uniquely mapped reads when using more complex annotations.

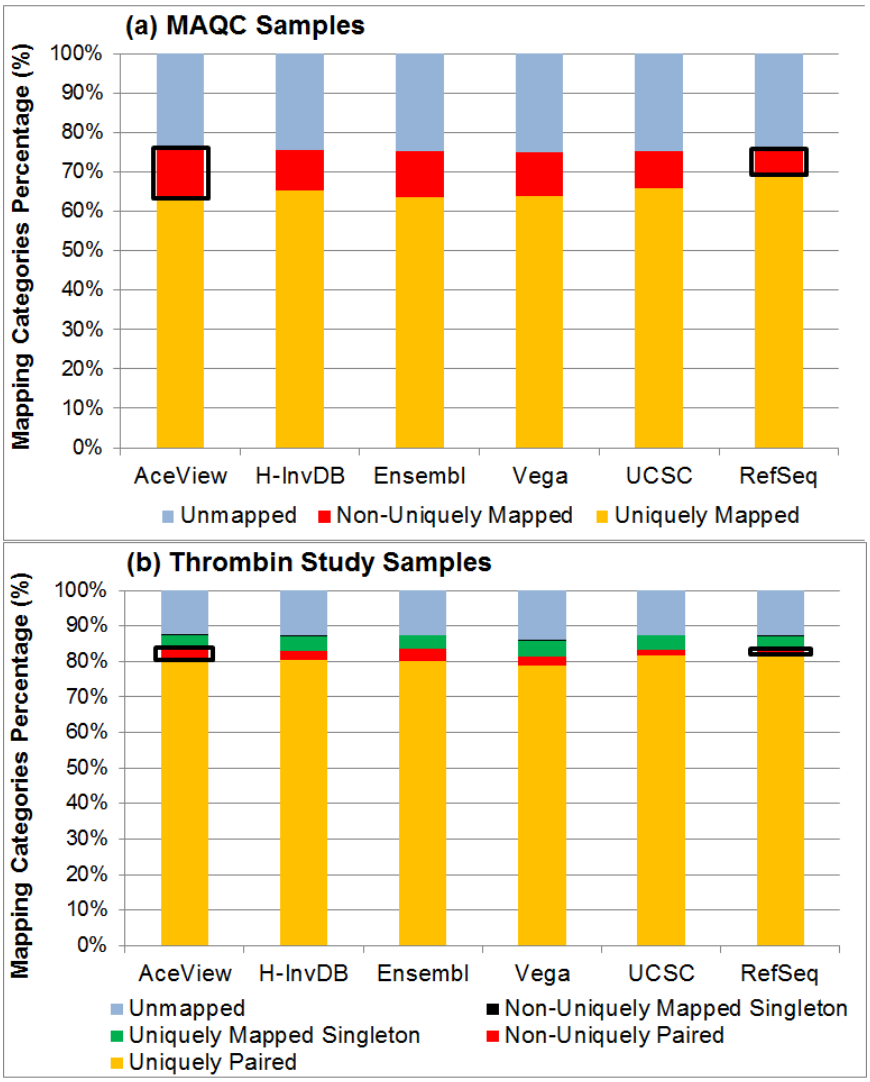


Figure 12: Distribution of Read Mapping Categories. (a) MAQC samples (SRA: SRP000727) contain single-end reads, and thus, there are three read mapping categories—uniquely mapped reads, non-uniquely mapped reads, and unmapped reads. (b) Thrombin study samples (SRA: SRP008482) contain paired-end reads, and thus, five read mapping categories can possibly occur. Cases of uniquely paired reads and non-uniquely paired reads occur when both ends of a read pair mapped to a reference genome. Situations of uniquely mapped singletons and non-uniquely mapped singletons occur when only one end of a read pair mapped to the reference genome. The RefSeq annotation has the highest percentage of uniquely mapped reads and the lowest non-uniquely mapped reads for both cases.

I then examined the percentage of reads that mapped to the annotated and un-annotated genomic sequences. More reads mapping to the annotated genomic sequences implies that more information will be available for the quantification step. From **Figure 13**, I observed that the AceView annotation resulted in the highest percentage of reads that map to annotated sequences. In contrast, the UCSC and RefSeq annotations had lower percentages of reads that map to annotated sequences, with UCSC being the lowest. Other than this outlying case, this evaluation metric followed the annotation complexity measure.

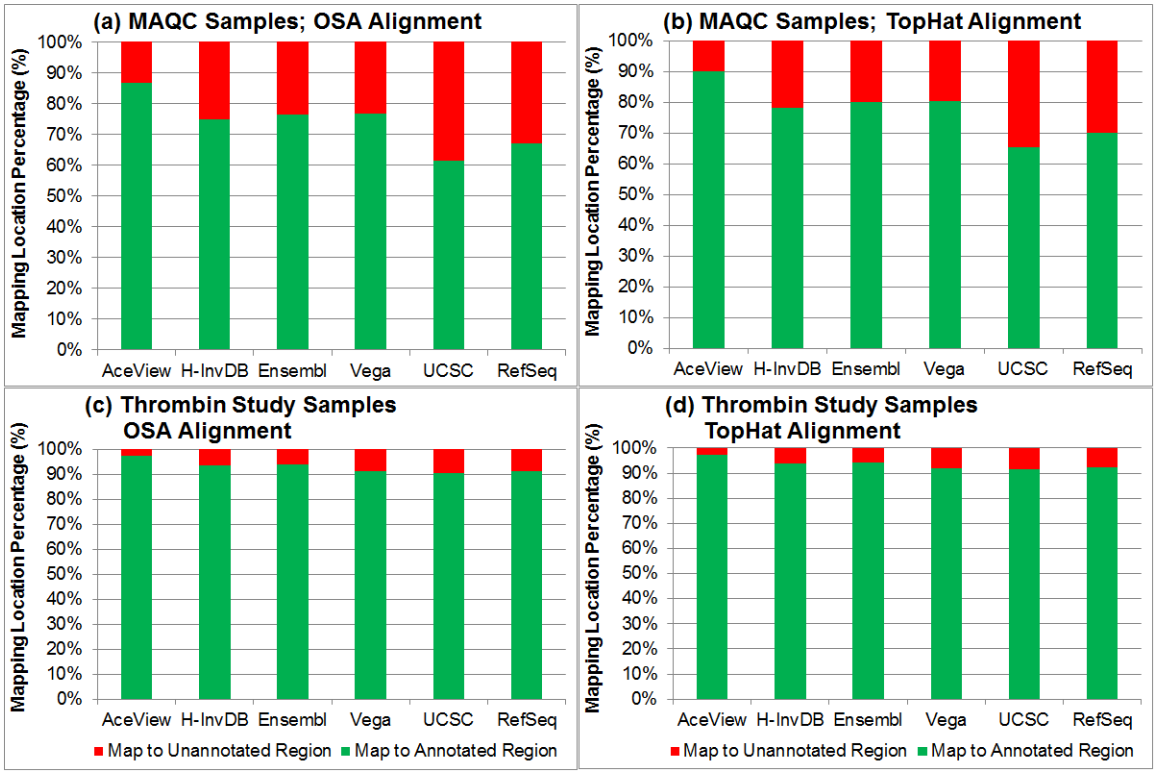


Figure 13: Percentage of Reads Mapping to Annotated/Unannotated Regions. Panels (a) – (d) represent different combinations of samples (top—MAQC samples; bottom—thrombin study samples) and spliced mappers (left—OSA; right—TopHat). The UCSC annotation usually has the lowest percentage of reads that mapped to the annotated genomic sequences, while the AceView annotation usually has the highest percentage. The same observation is applicable to all four combinations of samples and mappers.

Effect of human genome annotation complexity on quantification

I proposed two metrics to assess the impact of genome annotation complexity on RNA-seq quantification. The first metric was to evaluate the stability of gene and transcript expression estimates. **Figure 14** demonstrates the variation of average CoV due to the choice of the genome annotation and the selection of gene or transcript subgroups. I focused on four subgroups—all genes of each annotation, common genes (13,613 genes for the OSA pipeline and 13,810 genes for the TopHat-Cufflinks pipeline) that are defined in all annotations, genes not common to all annotations (i.e., uncommon genes), and all transcripts. Trends for all genes and uncommon genes were similar. The AceView annotation had the highest average CoV, followed by the Vega annotation, the Ensembl annotation, the H-InvDB annotation, the UCSC annotation, and the RefSeq annotation. In the case of all transcripts, sometimes the H-InvDB annotation resulted in the highest average CoV. For common genes, the difference in average CoV among various annotations was not significant. The RefSeq annotation always resulted in the lowest average CoV, whereas the H-InvDB or AceView annotations had the highest average CoV. The variation between annotations became larger for the cases of all genes, uncommon genes, and all transcripts since more annotation-specific elements were being considered. More complex annotations are more challenging for quantification because a larger number of ambiguous mappings occur. Note that Ensembl and Vega deviated from the trend of the annotation complexity measure. A possible rationale for this observation was that the Ensembl and Vega annotations tended to include more small RNAs compared with the other annotations. Since the sequencing data I analyzed follows the poly(A)-enrichment library preparation protocol, ideally, only mRNAs were retained in

the final sequencing libraries. Thus, the majority of small RNAs should have zero or very low expression. I defined these zero expressing elements as absent genomic elements. The inclusion of low-expressing genomic elements in the Ensembl or Vega annotation resulted in larger average CoV.

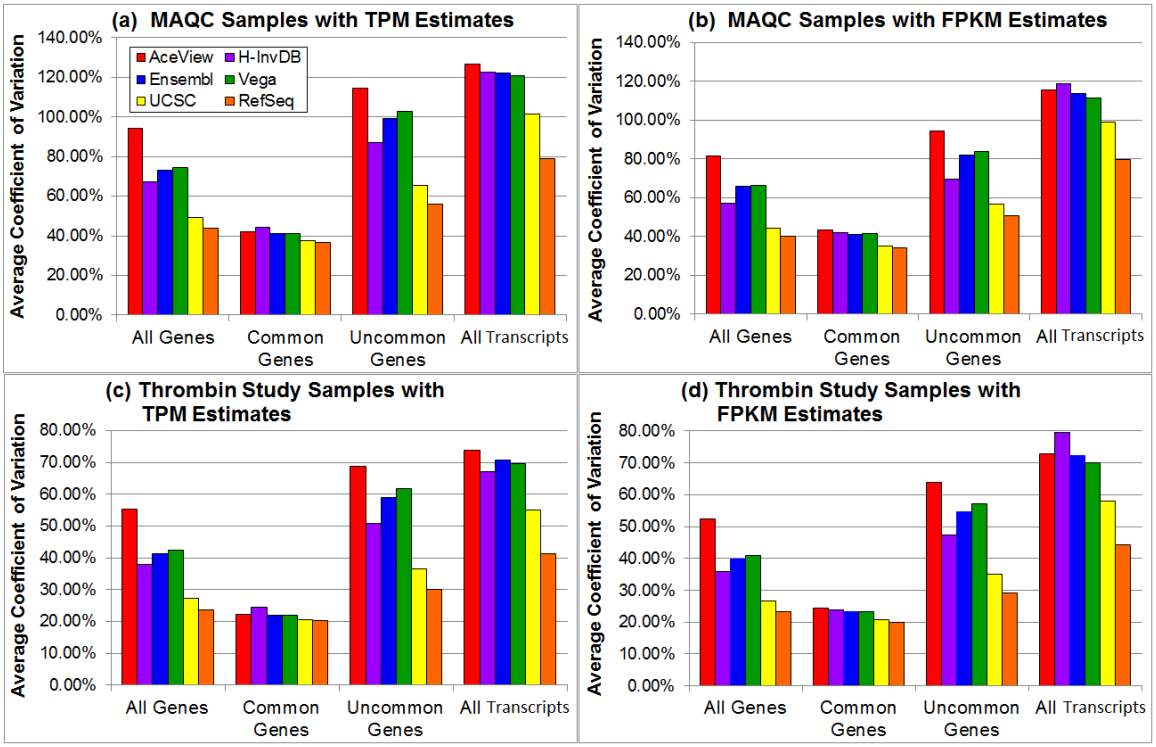


Figure 14: Average CoV for Various Annotations and Gene / Transcript Sets. Panels (a) – (d) represent different combinations of samples (top—MAQC samples; bottom—thrombin study samples) and expression estimates (left—TPM estimates from OSA package; right—FPKM estimates from TopHat alignment with Cufflinks quantification). The RefSeq annotation always has the smallest average CoV, while the AceView annotation has the highest average CoV for most of the cases. The variation is small when focusing on only common genes.

Figure 15 demonstrates that the percentage of present genomic elements depends on the annotation. I defined a "present" genomic element to be an element that has nonzero expression for at least one technical replicate. For common genes, all annotations had a similar percentage of present genes. For uncommon genes, all genes, and all transcripts, the relation among the AceView, H-InvDB, Ensembl, and Vega annotations was more uncertain compared with other evaluation metrics. In most cases, the H-InvDB annotation had a higher percentage of present genes / transcripts than the AceView annotation. The RefSeq annotation always had the highest percentage of present genes / transcripts, followed by the UCSC annotation. As I explained in the previous paragraph, more small RNAs are included in the Ensembl and Vega annotations. Because of the poly(A)-enrichment library preparation, most of these small RNAs had zero expression and were identified as absent, which correspondingly decreased the percentage of present genes or present transcripts.

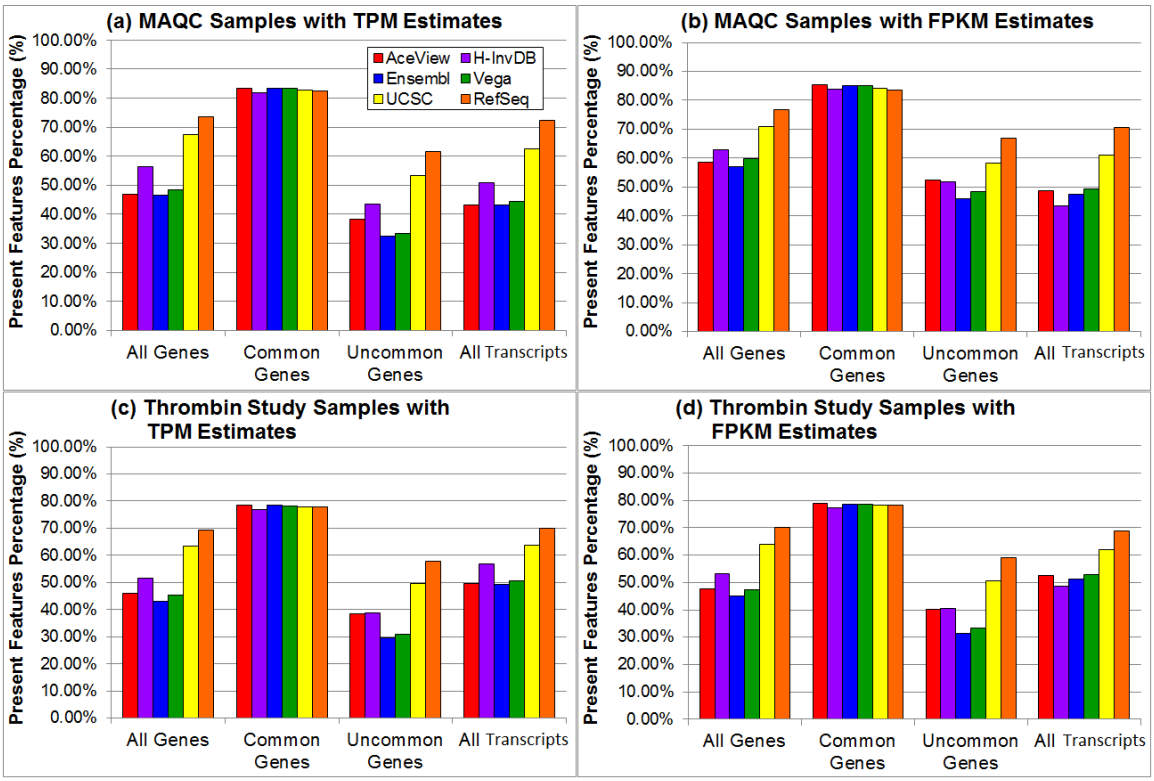


Figure 15: Present Percentage for Various Annotations and Gene / Transcript Sets. Panels (a) – (d) represent different combinations of samples (top—MAQC samples; bottom—thrombin study samples) and expression estimates (left—TPM estimates from OSA package; right—FPKM estimates from TopHat alignment with Cufflinks quantification). The RefSeq annotation usually has the highest percentage of present genomic elements, while the Ensembl or Vega annotation generally has the lowest percentage of present genomic elements. The variation is small when focusing on only common genes.

Effect of annotation complexity on differential expression calling

Three genes were validated by a qPCR assay for the thrombin study samples. I examined the difference between RNA-seq-based fold changes and qPCR-based fold changes and summarized the results in **Table 10**. From **Table 10**, I observed that the UCSC annotation always outperformed RefSeq annotation in terms of MAD from the qPCR fold-change estimates. However, the difference between them was not significant. In contrast, AceView and H-InvDB annotations had relatively higher MAD. Such the

observation leads to a conclusion that more complex annotations increase the difficulty of acquiring accurate gene expression estimates. Higher variations in gene expression estimates propagate to fold-change estimates.

Table 10: Comparison between qPCR-based and RNA-seq-based Fold Changes.

Gene	qPCR (FC)	RNA-seq TPM-Normalized Expression Estimates (FC)					
		AceView	***H-InvDB***	***Ensembl***	***Vega***	***UCSC***	***RefSeq***
TRAF1	2.862	3.029	3.034	3.025	2.998	2.934	2.922
FANCD2	-1.050	-0.782	-0.687	-0.888	-0.856	-0.840	-0.840
CELF1	-0.202	-0.138	-0.202	-0.098	-0.098	-0.239	-0.275
MAD between qPCR & RNA-seq		**0.166**	**0.178**	**0.143**	**0.145**	**0.106**	**0.114**
Gene	**qPCR (FC)**	**RNA-seq FPKM-Normalized Expression Estimates (FC)**					
		AceView	***H-InvDB***	***Ensembl***	***Vega***	***UCSC***	***RefSeq***
TRAF1	2.862	3.874	3.845	3.797	3.719	3.677	3.674
FANCD2	-1.050	0.057	0.057	-0.345	-0.322	-0.202	-0.151
CELF1	-0.202	0.642	0.516	0.595	0.585	0.390	0.356
MAD between qPCR & RNA-seq		**0.987**	**0.936**	**0.812**	**0.791**	**0.751**	**0.756**

Fold changes (FC) are defined as the ratio of the average expression of thrombin-treated samples to that of control samples.

The qPCR data for MAQC samples are publicly available. I used three statistics to assess variations due to the genome annotation choice. As shown in **Figure 16**, less complex genome annotations (e.g., the RefSeq annotation) result in lower MAD, lower RMSE, and higher correlation coefficients when comparing RNA-seq fold-change estimates to qPCR fold-change estimates. Some outlying cases existed (e.g., the Ensembl annotation had the highest RMSE when using FPKM expression estimates), but the general trend of this evaluation metric still followed the annotation complexity measure.

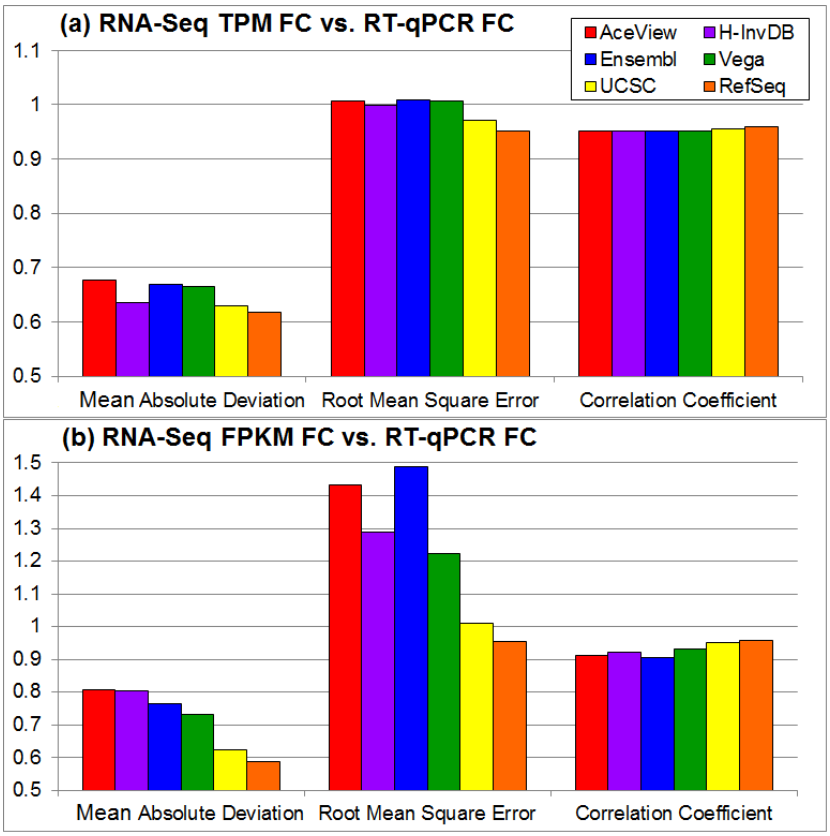


Figure 16: Statistics for Fold-Change Comparisons. The comparison of fold-change estimates between RNA-seq and qPCR using two RNA-seq expression estimates and three statistics. (a) TPM estimates are produced by the OSA package. (b) FPKM estimates are generated by Cufflinks with TopHat alignment. The RefSeq annotation always has the lowest mean absolute deviation, the lowest root-mean-square error, and the highest correlation coefficient when treating qPCR estimates as the ground truth.

2.4.1.5 Summary of Case Study

The genome annotation is a necessary component for RNA-seq expression analysis. Multiple genome annotations are publicly available; however, it is not clear how different choices of the genome annotation will affect downstream RNA-seq expression estimates. In this case study, I defined the complexity of the human genome annotation and assessed the relationship between genome annotation complexity and several RNA-seq performance metrics. Based on my complexity measure, I ordered existing human

genome annotations from most to least complex as follows—AceView, H-InvDB, Ensembl, Vega, UCSC, and RefSeq. In more complex annotations, a higher percentage of the entire genome is annotated. For RNA-seq sequence mapping, less complex annotations resulted in a higher percentage of uniquely mapped reads and uniquely mapped read pairs for both single-end and paired-end samples. However, at the same time, the number of RNA-seq reads mapping to annotated genomic sequences was smaller for less complex annotations. Genome annotation complexity also affected RNA-seq expression estimates. More complex annotations resulted in more ambiguous mappings, which increased the difficulty of RNA-seq quantification and caused higher expression variation among RNA-seq replicate libraries. Furthermore, more complex annotations led to a lower percentage of present (i.e., detected) genes or transcripts, which suggests that the putative genomic elements in these annotations tend to be non-expressers or low expressers. Deviations in RNA-seq expression estimates due to differences in genome annotation complexity can propagate to fold-change statistics and, subsequently, differential expression detection. When comparing RNA-seq fold-change statistics to ground-truth qPCR fold-change statistics, more complex annotations tended to have larger deviation and smaller correlation. In summary, the impact of genome annotation choice on RNA-seq expression estimates is significant, and the choice of annotation should depend on the objective of an application. Less complex genome annotations are preferable for applications that require more stable RNA-seq expression estimates. However, to discover and explain unknown biological mechanisms, more comprehensive and complex genome annotations may be necessary.

2.4.2 Impact of Expression Quantification Choice on Feature Quality

2.4.2.1 Background

A single RNA-seq run can produce millions of sequence reads. Focusing on RNA-seq transcriptome profiling, a standard data analysis pipeline includes sequence mapping, expression quantification and normalization, and various downstream inferences. Expression quantification algorithms attempt to uniquely assign sequence reads to genes or transcripts. However, this is a challenging process since (1) alternative spliced transcripts of a gene share exons and (2) sequence reads may map to multiple loci due to the relatively short read length and high similarity among some genomic regions [72]. These challenges result in read assignment uncertainty.

To address these challenges, researchers have developed a number of quantification algorithms. Several (e.g., HTSeq [44] and BEDTools [45]) simplify the problem by counting the number of sequence reads aligned to a targeted gene with a predefined gene model, and others (e.g., Cufflinks [47] and RSEM [46]), which potentially resolve the multi-mapping issue, build upon the Poisson-based model and probabilistically assign sequence reads to transcripts or genes. Quantification algorithms can be classified into two categories in terms of the input information from the sequence alignment. Some require sequence alignment to be reported in genomic coordinates while others require sequence alignment to be reported in transcriptomic coordinates.

Because of various characteristics of quantification algorithms, I proposed an approach for systematically assessing the performance of these algorithms using both simulated and real datasets.

2.4.2.2 Experimental Design

The workflow of this case study is shown in **Figure 17**. Data sources included one simulated dataset and two real datasets. I mapped sequence reads of these datasets to the UCSC hg19 reference genome using TopHat [83] either with or without external genome annotation information (the GTF file) and to the RefSeq reference transcriptome using Bowtie [73]. I then used Cufflinks [47], HTSeq [44], and MISO [88] to quantify sequence alignments reported in genomic coordinates, and RSEM [46], eXpress [139], and MMSEQ [140] to quantify those reported in transcriptomic coordinates. Finally, I computed gene and transcript read counts from the simulated dataset as the ground truth, and investigated the performance of various quantification algorithms.

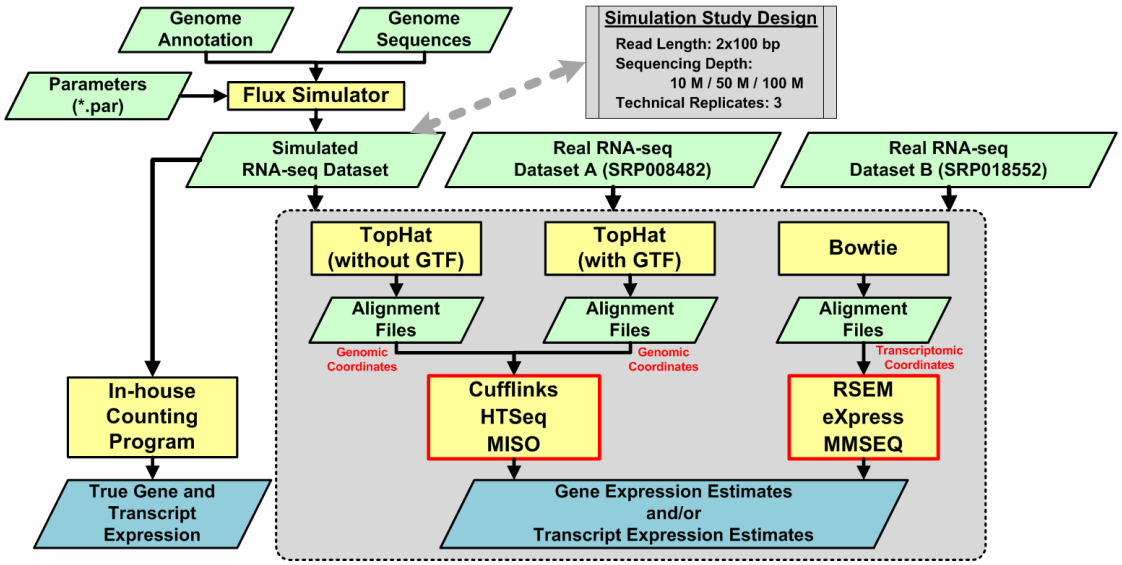


Figure 17: Workflow for Chapter 2, Case Study 2. This case study includes three RNA-seq datasets—one simulated and two real. Depending on sequence mapping outputs, some quantifiers (e.g., Cufflinks, HTSeq, and MISO) are designed to handle mapping outputs in genomic coordinates, while some others (e.g., RSEM, eXpress, and MMSEQ) are designed to handle mapping outputs in transcriptomic coordinates. With true expression derived from the simulated dataset and estimated expression generated by the six quantifiers for the three datasets, I designed various metrics to assess the performance of each quantification pipeline.

2.4.2.3 Datasets

To create the simulated dataset, I used Flux Simulator [141] to produce a gene expression profile and then used the same expression profile to generate three technical replicates with the read length of 2×100 bp for each sequencing depth (i.e., 10, 50, and 100 million read pairs). In Flux Simulator, multiple library preparation and sequencing steps introduced variations among the technical replicates.

I also downloaded two publicly available datasets from the NCBI SRA repository. The first dataset, which contains three thrombin-treated samples and two control samples, studied the effect of thrombin on endothelial function (SRA accession: SRP008482 [137]). Using the Illumina HiSeq 2000 platform, the authors sequenced these samples, each of which has around 50 million 2×100 bp read pairs. The second dataset, which contains four treatments, each with two replicates, investigated the off-target effect of EGFP siRNA and pro-siRNA in the HeLa-d1EGFP cell line (SRA accession: SRP018552 [142]). Using the Illumina Genome Analyzer II platform, the authors sequenced these samples, each with about 25 million 50 bp single-ended reads.

2.4.2.4 Results and Discussion

For this case study, I proposed three assessment metrics to investigate the performance of quantification algorithms. **Figure 18**, panel A and **Figure 19**, panel A show that normalized RMSE varies according to the quantification strategy and the sequencing depth. A lower normalized RMSE indicates smaller deviation between estimated and true expression. Note that the high sequencing depth (e.g., 100M read pairs) does not improve significantly in the normalized RMSE. **Figure 18**, panel B and **Figure 19**, panel B demonstrate that the number of genes / transcripts falling within

predefined percent errors of the CoV differs according to the quantification strategy. Curves closer to the upper-left corner of the figure indicate a closer measure of variation to the ground truth. **Figure 20** and **Figure 21** use box plots to illustrate that the distribution of the gene-wise or transcript-wise CoV also varies according to the quantification strategy in the real RNA-seq datasets. Lower CoV indicates a smaller variation among the technical replicates, which is the desired property.

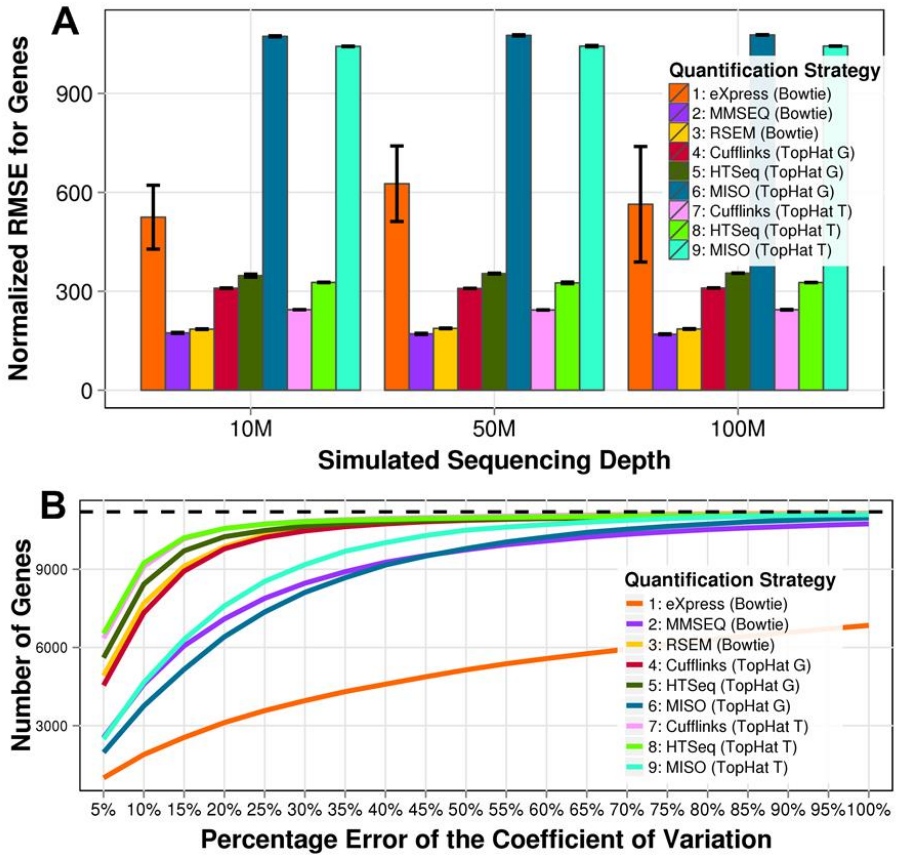


Figure 18: Gene Expression Quantification Performance for the Simulated Data. Gene-level performance metrics for the simulated RNA-seq dataset includes (A) the normalized RMSE and (B) the percentage error of the CoV. TopHat T denotes genome alignment using TopHat with external GTF information; TopHat G denotes genome alignment using TopHat without external GTF information; and Bowtie represents transcriptome alignment using Bowtie.

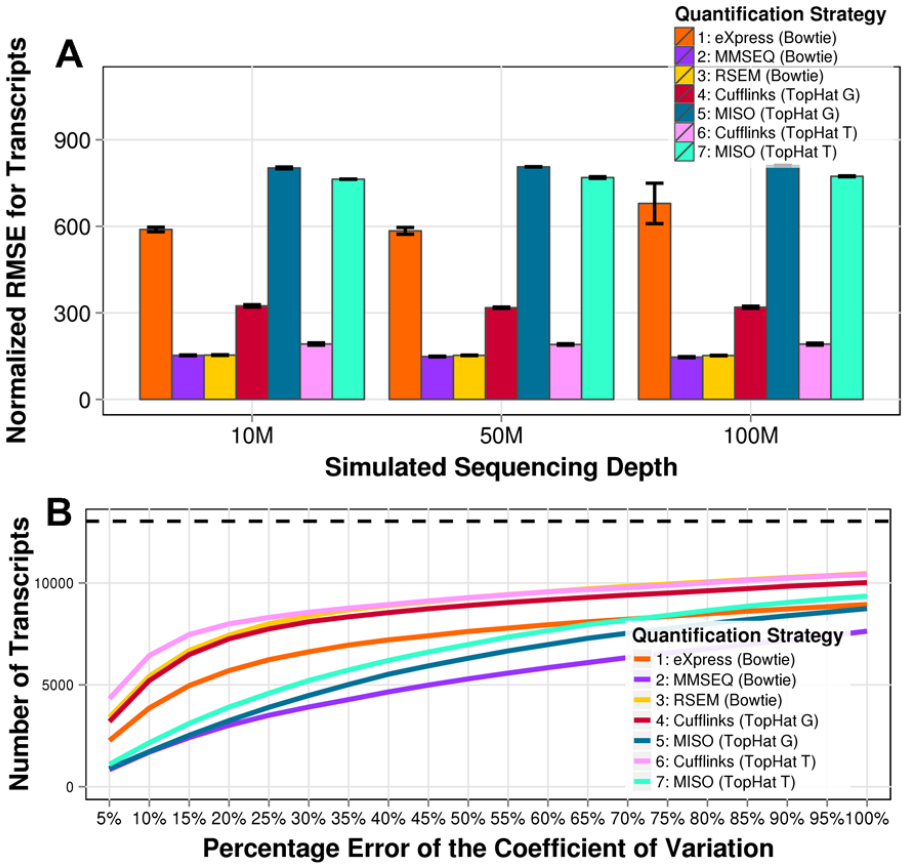


Figure 19: Transcript Expression Quantification Performance for the Simulated Data. Transcript-level performance metrics for the simulated RNA-seq dataset includes (A) the normalized RMSE and (B) the percentage error of the CoV. TopHat T denotes genome alignment using TopHat with external GTF information; TopHat G denotes genome alignment using TopHat without external GTF information; and Bowtie represents transcriptome alignment using Bowtie.

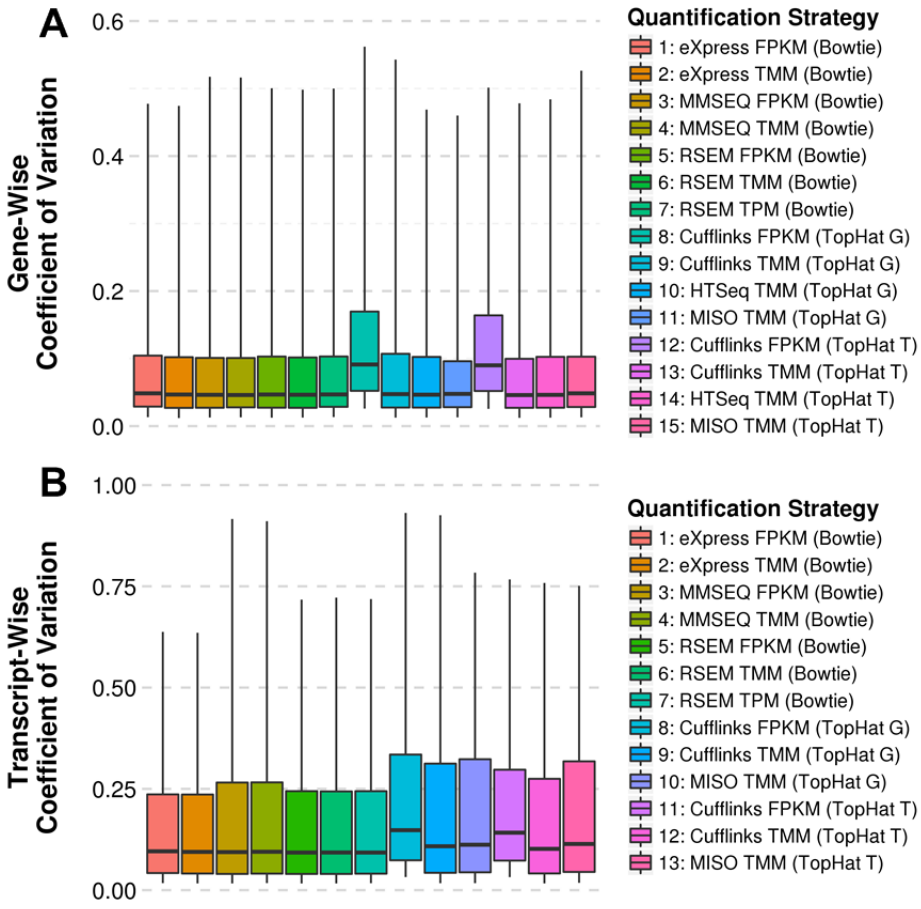


Figure 20: Expression Quantification Performance for the SRP008482 Data. The distribution of the CoV varies according to the quantification strategy for both gene- and transcript-level expression estimates using the SRP008482 dataset. TopHat T denotes genome alignment using TopHat with external GTF information; TopHat G denotes genome alignment using TopHat without external GTF information; and Bowtie represents transcriptome alignment using Bowtie.

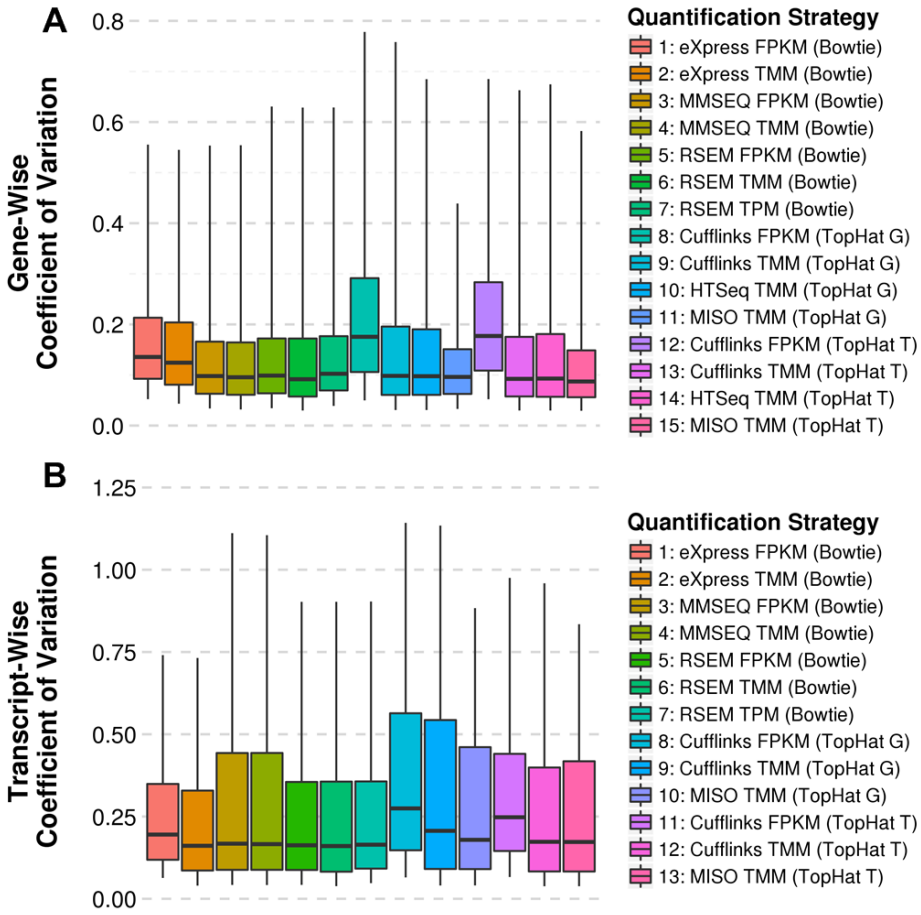


Figure 21: Expression Quantification Performance for the SRP018552 Data. The distribution of the CoV varies according to the quantification strategy for both gene- and transcript-level expression estimates using the SRP018552 dataset. TopHat T denotes genome alignment using TopHat with external GTF information; TopHat G denotes genome alignment using TopHat without external GTF information; and Bowtie represents transcriptome alignment using Bowtie.

My results suggest that genome alignment with external GTF information (denoted as TopHat T) resulted in smaller deviations and lower CoVs than that without GTF information (denoted as TopHat G). This observation held for Cufflinks, HTSeq, and MISO with the three metrics, shown in **Figure 18**, **Figure 19**, **Figure 20**, and **Figure 21**. Without external GTF information, TopHat searches for candidate mapping locations in the entire genome, which may increase the odds that a sequence read maps to an incorrect location. When focusing on expression-based applications, if the GTF file is available, I recommend using GTF-guided alignment to speed up the computational process and increase the expression estimation accuracy.

Another key finding was that quantification algorithms based on transcriptome alignment tended to result in smaller deviations than genome alignment when using simulated expression as the reference, with eXpress being an outlier (**Figure 18**, panel A and **Figure 19**, panel A). The analysis of real RNA-seq datasets showed that transcriptome alignment also resulted in a comparable or lower CoV across technical replicates than genome alignment (**Figure 20** and **Figure 21**). The difference became more significant in the transcript-wise CoV analysis. However, for the percent error of the CoV, quantification algorithms based on genome alignment, such as Cufflinks and HTSeq, outperformed all other quantification algorithms, with RSEM being tied with the Cufflinks TopHat G quantification strategy (**Figure 18**, panel B and **Figure 19**, panel B). Based on these observations, I infer that quantification algorithms based on genome alignment may cause greater absolute deviation but maintain relative variation among technical replicates. Even though the observations were not exactly consistent, quantification algorithms based on transcriptome alignment generally performed better in

terms of the three metrics, with RSEM outperforming others in all cases. The possible reason is that transcriptome alignment preserves direct mapping information about each transcript while genome alignment does not. Thus, while genome alignment requires additional effort to assign reads to one of the transcripts of a gene, transcriptome alignment requires only the identification of chimeric mappings in multi-mapping cases.

My study identified several outlying cases. For one, with three evaluation metrics, the MISO package performed worse (i.e., a greater deviation and a higher CoV) in most of the cases, which may have been due to MISO ignored too much information, such as (1) read pairs mapped to the same strand and (2) reads that have no paired mate. Moreover, Bayesian-based algorithms (e.g., MISO and eXpress) tended to introduce higher variation in expression estimates. I hypothesized that Bayesian-based approaches heavily depend on the prior distribution. Thus, if algorithms cannot converge within a predefined number of iterations, the expression estimates may have been only a suboptimal solution. This hypothesis was partially confirmed by the high sensitivity of eXpress to the "forgetting factors" discussed in [139].

2.4.2.5 Summary of Case Study

I proposed an approach that includes three alignment strategies and six quantification algorithms for assessing RNA-seq quantification algorithms in replication studies using both simulated and real RNA-seq datasets. By examining multiple metrics, I found that the TopHat T alignment strategy always outperformed the TopHat G alignment strategy. In addition, quantification algorithms using sequence alignment reported in transcriptomic coordinates usually resulted in a smaller deviation and a lower CV, with eXpress being the outlier. Furthermore, RSEM consistently performed better

compared with other quantification algorithms. My approach is useful for comprehensively assessing new RNA-seq quantification algorithms. Based on my results, I suggest using quantification algorithms with transcriptome alignment. If only genome alignment is possible and external GTF information is available, incorporating GTF information will yield higher performance.

2.4.3 Impact of Expression Normalization Choice on Feature Quality

2.4.3.1 Background

RNA-seq for quantifying gene or transcript expression, one of the major applications of the NGS technology, has received increased attention because of its potential to replace the microarray technology. Some of the perceived benefits of RNA-seq over microarrays include (1) improved dynamic range of expression detection and (2) the ability to detect a wide variety of RNA forms (e.g., small RNAs and splice variants) [32, 135]. Analogous to microarrays, normalization of RNA-seq data to obtain quantitative and comparable gene or transcript expression values is an important step [91, 143, 144]. Several experimental factors in the sequencing process such as library preparation, sequencing depths, and base calling methods can introduce biases in downstream RNA-seq analysis. The purpose of the normalization step is to detect and adjust such biases. However, it is unclear how existing RNA-seq normalization methods handle various gene expression distributions. Using a simulated dataset, I compared several existing methods for RNA-seq expression normalization and evaluated them in terms of the recovery of designed fold changes.

Most RNA-seq expression normalization methods are simple global normalization techniques that use a constant scaling factor for each sequencing sample. In this case

study, I investigated four existing RNA-seq normalization methods, including RPM / FPM [90], TMM [97], RLE [98], and Upper Quartile [96]. RPM / FPM adjust the total number of mapped reads per sample. However, RPM / FPM can be biased by relatively small proportions of highly-expressed genes and, as such, can bias DEG detection [96]. The number of reads expected to map to a gene is not only dependent on the expression level and the length of the gene, but also on the composition of the sampled RNA population. Normalization procedures such as TMM, RLE, and Upper Quartile attempt to estimate scaling factors between two samples to adjust total RNA output [97]. The TMM method trims log-ratio (M values) and log-average (A values) to find possible sets of stably expressed genes to estimate scaling factors. The RLE method generates a reference library by calculating the geometric mean of each gene across all samples, and the median ratio of each sample to the reference is taken as the scaling factor. The Upper Quartile method uses the ratio of the upper quartile between two samples as the scaling factor. The TMM, RLE, and Upper Quartile normalization also belong to global normalization methods. The difference between these methods and RPM / FPM is that they consider adjusting the total RNA output rather than the library size, which can reduce biases caused by highly-expressed genes.

2.4.3.2 Datasets

I assessed the robustness of normalization methods to various fold-change distributions using a simulated RNA-seq dataset. **Table 11** summarizes five simulated gene expression distributions. I directly simulated raw counts of gene expression so as to eliminate possible errors introduced from the sequencing and sequence mapping steps. Using this dataset, I aim to investigate pros and cons of each normalization method.

Table 11: Simulated RNA-seq Expression Distribution.

Distribution	Description
Simulated—Reference (Data T)	Uniformly sample 200 gene expression from the sorted RNA-seq gene expression data
Simulated—Variant 1 (Data Var1)	Randomly select 5% of the 200 genes in "Simulated-Ground Truth" and increase their expression by 10-fold
Simulated—Variant 2 (Data Var2)	Randomly select 20% of the 200 genes in "Simulated-Ground Truth" and increase their expression by 3-fold; and then randomly select another non-overlapping 20% of the 200 genes and decrease their expression by 3-fold
Simulated—Variant 3 (Data Var3)	Uniform expression for all the 200 genes with gene expression equals to the median expression in "Simulated-Ground Truth"
Simulated—Variant 4 (Data Var4)	70% lowest expressing genes in "Simulated-Ground Truth" are set to be 0 expression

2.4.3.3 Results and Discussion

In **Figure 22**, panel (a), I computed fold changes between Data Var1 and Data T and expected a fold change of 10 for 10 genes. RLE, TMM, and Upper Quartile methods were able to recover fold changes with less than 10% errors, but RPM was not. In **Figure 22**, panel (b), I computed fold changes between Data Var2 and Data T. In this case, RLE and TMM were able to recover fold changes with less than 10% errors, whereas RPM and Upper Quartile were not. **Figure 22**, panel (c) shows results for an extreme, unrealistic case, that is, gene expression is uniformly distributed (Data Var3). TMM was the only method that correctly recovered fold changes with less than 10% errors. **Figure 22**, panel (d) demonstrates results for another less realistic case in which the data contains excessive zero-expressing genes (Data Var4). RPM and TMM failed in this case.

These results suggested that TMM and RLE methods are more robust expression normalization methods because they were capable of recovering fold changes within the

10% error margin for various simulated fold-change distributions. The RPM and Upper Quartile methods were susceptible to fold-change distributions with either a few highly-expressed genes or many differentially expressed genes.

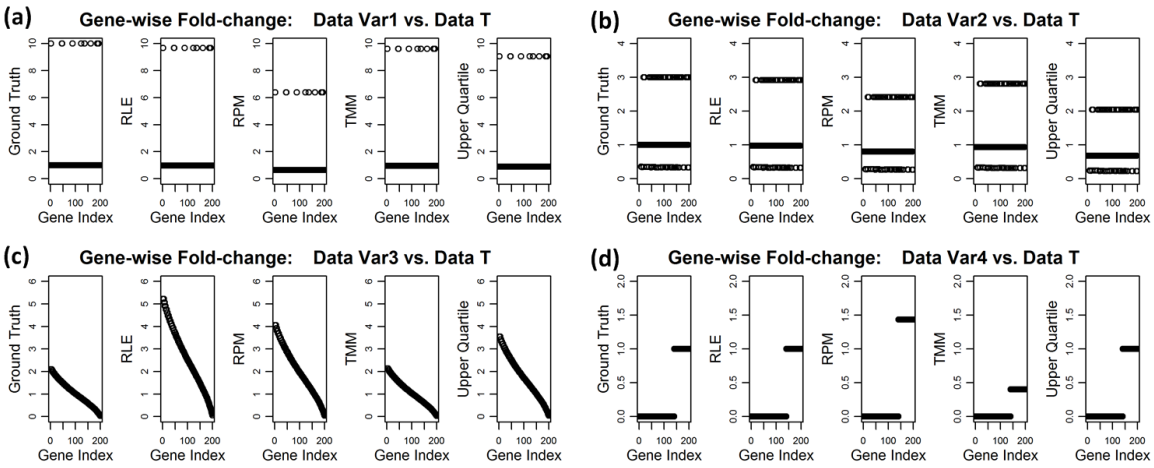


Figure 22: Recovered Fold Changes for Various Simulated Distributions. Panels (a) – (d) demonstrate the performance of several RNA-seq normalization methods for various simulated fold-change distributions. Each panel consists of five plots. From left to right, they are ground-truth fold changes, fold changes derived from RLE-normalized expression, RPM-normalized expression, TMM-normalized expression, and Upper Quartile-normalized expression.

2.4.3.4 Summary of Case Study

In this case study, I explored some existing RNA-seq normalization methods, including RPM, TMM, RLE, and Upper Quartile, and assessed their capability of tolerating different RNA-seq DEG distributions. Using simulated RNA-seq fold-change distributions, I observed that TMM and RLE failed only in one of the extreme scenarios. Upper Quartile and RPM were both sensitive to the distribution of fold changes. Thus, fold-change (or DEG) distribution is an important factor when choosing a normalization method for RNA-seq expression analysis.

2.4.4 Impact of Pipeline Choice on Feature Quality

2.4.4.1 Background

The first phase of the FDA-led microarray quality control project (MAQC-I) investigated and compared different genomic microarray technologies for quality verification [145]. The second phase of the project, MAQC-II, studied 30,000+ microarray gene expression-based data analysis pipelines to assess their prediction reproducibility for regulatory purposes [146]. Based on MAQC-I and MAQC-II, the FDA initiated MAQC-III (also known as sequencing quality control [SEQC]), which was an in-depth assessment of RNA-seq [32, 90, 135, 147]. Specifically, the goal of SEQC was to conduct a comprehensive evaluation of both RNA-seq technology and RNA-seq data analysis pipelines (i.e., RNA-seq pipelines), which was similar to the MAQC-I and MAQC-II evaluation of microarrays, as part of the FDA Critical Path Initiative (http://www.fda.gov/oc/initiatives/criticalpath/). While Su *et al*. summarized the SEQC RNA-seq technology investigation [148], this complementary case study summarizes the RNA-seq pipeline investigation. In particular, this case study examines the effect of RNA-seq pipelines on gene expression accuracy, precision, reliability, and reproducibility (defined in Section 2.3).

For biological and medical applications, choosing a proper pipeline for RNA-seq gene expression remains a critical challenge due to its relative immaturity (i.e., fewer standards reported compared to microarrays), complexity, and diverse applicability. We performed a literature survey on RNA-seq pipelines consisting of sequence mapping [38, 39, 73, 75, 82, 84, 85, 128, 129, 131], expression quantification [44, 46, 47, 72], and expression normalization [90, 97-99]. Multiple comparative investigations exist for

individual components of RNA-seq pipelines—mapping [128, 149-152], quantification [46, 95, 153, 154], and normalization [99, 155, 156]. However, despite the interdependence of these components [72], their joint impact is seldom comprehensively investigated. A previous investigation of 50 RNA-seq pipelines examined combinations of ten mapping and five quantification algorithms, but did not consider the interactive effects of different normalization algorithms [113]. Another study investigated three mapping, two quantification, and five DEG detection methods to assess concordance of gene expression and DEGs, but did not report on effects of interactions among pipeline components [112]. Both of these studies did not consider the effect of pipeline choice on downstream applications such as gene expression-based prediction. To the best of our knowledge, *no studies have reported the joint effects of all components in RNA-seq pipelines and no guidelines exist for selecting RNA-seq pipelines for downstream prediction of disease outcome.*

The FDA first coordinated multiple sites of SEQC to generate a multi-replicate benchmark dataset (referred to as SEQC-Benchmark in this case study) [148], and then provided the dataset to our team to investigate the joint impact of pipeline components on gene expression estimation. Our team developed evaluation metrics (i.e., accuracy, precision, reproducibility, and reliability) for assessing a representative set of 278 RNA-seq pipelines using the SEQC-benchmark dataset.

2.4.4.2 Experimental Design

We systematically investigated 278 RNA-seq pipelines (**Table 12**) that included combinations of 13 sequence mapping algorithms (**Table 6**) [38, 39, 73, 75, 82-84, 128, 129, 131], three categories of expression quantification algorithms (**Table 7**) [44, 46, 47],

and seven expression normalization methods (**Table 8**). Sequence mapping algorithms were further categorized based on mapping strategy and mapping reporting. Mapping strategy refers to spliced or un-spliced algorithms. Un-spliced algorithms map whole read sequences while spliced algorithms split reads into smaller segments in order to accommodate long gaps such as introns. Mapping reporting refers to the number of mapping locations reported per read, either single-hit (i.e., one location reported per read) or multi-hit (i.e., multiple locations reported per read). To gain insight into these pipelines, we used the multi-site and multi-replicate SEQC-benchmark dataset along with a ground-truth quantitative PCR (qPCR) dataset. **Table 13** and **Table 14** summarize the SEQC-benchmark datasets and samples. Although qPCR can be variable and significantly different among platforms [148], we used it as a benchmark reference only after filtering the qPCR data based on the embedded ground-truth (**Figure 52**). The filtering process is detailed in the Appendix B section "Filtering the qPCR Benchmark Dataset to Produce a Reference Set of Genes."

Table 12: RNA-seq Pipelines Investigated in Case Study 4.

RNA-seq Pipeline Factors		# of Algorithm Choices	Algorithm Choices
Mapping	Algorithm	13	Bowtie, Bowtie2, BWA, GSNAP, Magic, MapSplice, Novoalign, OSA, RUM, STAR, Subread, TopHat, WHAM
	Strategy	2	Un-Spliced, Spliced
	Reporting	2	Single-Hit, Multi-Hit
Quantification		3	Count-Based, Cufflinks, RSEM
Normalization		7	FPM, FPKM, Median, Upper Quartile, RLE, TMM, Magic Expression Index
Total Pipelines		**278***	

*All combinations of mapping, quantification, and normalization are not possible because of incompatibility

Table 13: SEQC Benchmark Datasets.

Platform	Data Acquisition Site(s)	Samples Acquired	Replicates per Sample	Notes
Illumina HiSeq 2000	Beijing Genomics Institute (BGI) and Mayo Clinic (MAY)	A, B, C, D	4	Each sample replicate was sequenced in 16 lanes across two flow cells, but we used data from only two lanes of a single flow cell for this study.
Bio-Rad PrimePCR	FDA	A, B, C, D	1	The assay contains 20,801 genes.

Table 14: SEQC Benchmark Samples.

Sample Name	Sample Description
A	Stratagene's Universal Human Reference RNA (UHRR)
B	Ambion's Human Brain Reference RNA (HBRR)
C	A mixture of 75% A and 25% B
D	A mixture of 25% A and 75% B

As summarized in **Table 12**, the 13 mapping algorithms tested are Bowtie [73], Bowtie2 [75], BWA [38], GSNAP [82], Magic (a new pipeline developed by NCBI for the SEQC project) [121, 130], MapSplice [84], Novoalign (a commercialized package developed by Novocraft) [76], OSA [85], RUM [128], STAR [39], Subread [131], TopHat [83], and WHAM [129]. Some use un-spliced mapping of reads to the transcriptome, some others perform spliced mapping to the genome. The Magic pipeline uses both in parallel and compares the quality of each alignment to keep the best across multiple targets. Mapping algorithms may report only unique mapping, or allow for multiple mapping locations per read. Quantification algorithms include simple count-based methods (i.e., HTSeq [44]) and Poisson model-based probabilistic methods applied

to either genomic (i.e., Cufflinks [47]) or transcriptomic mapping data (i.e., RSEM [46]). The Magic, RUM, and Subread (i.e., featureCounts [132]) pipelines include embedded quantification methods that fall into the category of simple count-based methods. Normalization methods include simpler scaling methods (i.e., fragments per million mapped fragments [FPM], fragments per kilobase per million mapped fragments [FPKM], median, and upper quartile), more robust scaling methods (i.e., relative log expression [RLE] and trimmed mean of m-values [TMM]), and methods embedded in specific pipelines (i.e., Magic expression index).

2.4.4.3 Datasets

The FDA SEQC-benchmark dataset (Gene Expression Omnibus accession number GSE47792) includes paired-end RNA-seq data generated using the Illumina HiSeq 2000 platform with the read length of 100 bp [148]. We used a subset of the SEQC-benchmark dataset sequenced at two sites: Beijing Genomics Institute (BGI) and Mayo Clinic (MAY). We used four samples (i.e., A, B, C, and D), each with four replicate libraries prepared at the sequencing sites. Sample A contains the Universal Human Reference RNA (UHRR), sample B contains the Human Brain Reference RNA (HBRR), sample C contains a mixture of A and B (75% A and 25% B), and sample D contains a mixture of A and B (25% A and 75% B). We used data from two lanes of a single flow cell for each sample replicate. The SEQC also provided the benchmark qPCR dataset that includes 20,801 genes assayed with PrimePCR (Bio-Rad, Hercules, California). Each PrimePCR gene was assayed once for each of the four samples (i.e., A, B, C, and D). The FDA SEQC benchmark datasets and samples are summarized in **Table 13** and **Table 14**.

2.4.4.4 Results and Discussion

We systematically investigated 278 RNA-seq pipelines (**Table 12**) that included combinations of mapping, quantification, and normalization components listed in **Table 6**, **Table 7**, and **Table 8.** To gain insight into these pipelines, we used the multi-site and multi-replicate SEQC-benchmark dataset and the qPCR benchmark dataset. **Table 13** and **Table 14** summarize SEQC benchmark samples and datasets. **Table 15** summarizes the four benchmark metrics, including accuracy, precision, reliability, and reproducibility, used to evaluate pipelines. Details of these metrics have been discussed in Section 2.3.2.

Table 15: RNA-seq Pipeline Metrics for Case Study 4.

Metric	Description	Number of Raw Metric Values	Summary Statistics
Accuracy	Accuracy is defined as the deviation of RNA-seq pipeline-derived log ratios from the corresponding qPCR-based log ratios.	10,222 or 2,044	Median
Precision	Precision is defined as the coefficient of variation over sample replicate libraries	40,888 or 8,176 (10,222 or 2,044 genes × 4 samples)	
Reliability	Reliability is defined as the intraclass (or intra-sample in our case) correlation that quantifies how similar replicate libraries of a sample are to one another using ANOVA techniques	10,222 or 2,044	
Reproducibility	Reproducibility is defined as the Spearman correlation between two replicate libraries of the same sample.	24 (6 comparisons per sample × 4 samples)	

*All metrics are computed based on either 10,222 genes (denoted as "All Genes") or 2,044 genes (denoted as "Low-Expressing Genes")

Impact of mapping, quantification, and normalization on gene expression accuracy

We defined the accuracy metric as the deviation of RNA-seq pipeline-derived log ratios of gene expression from the corresponding qPCR-based log ratios, and visualized the median accuracy of all genes and low-expressing genes (refer to the Appendix B, Section "Filtering the qPCR Benchmark Dataset to Produce a Reference Set of Genes" for the definition of these gene sets) using heatmaps (**Figure 23**).

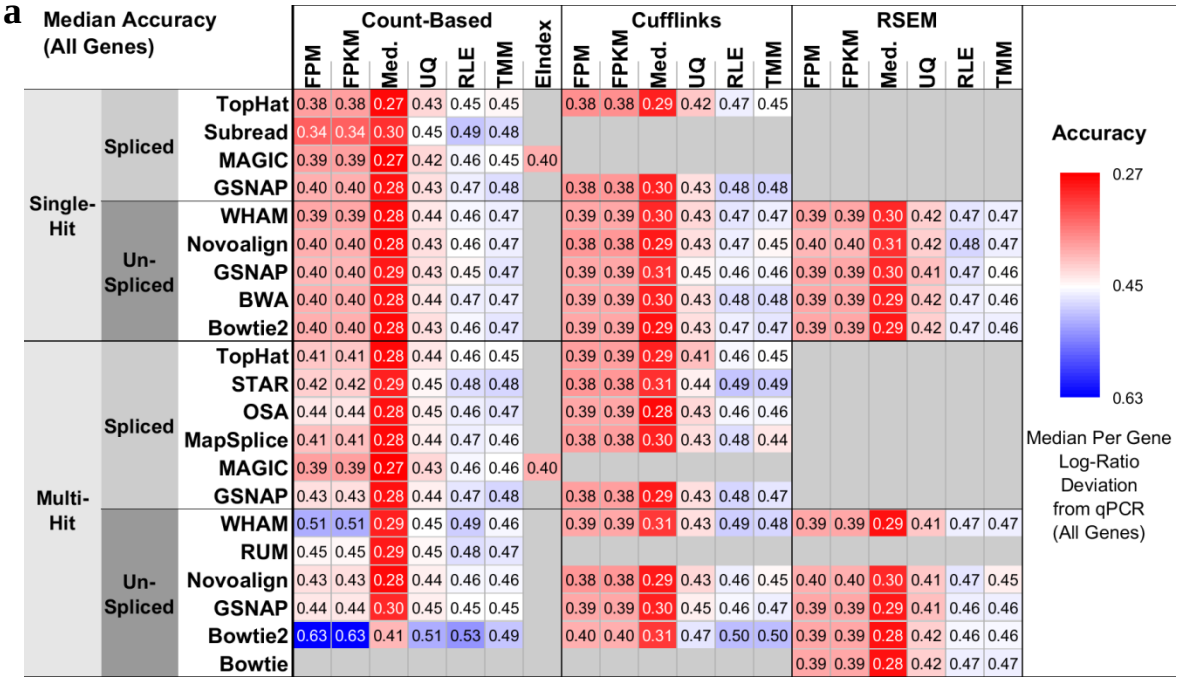


Figure 23: Median Accuracy of All and Low-Expressing Genes. The 278 RNA-seq pipelines applied to the SEQC-benchmark dataset differ in terms of the median accuracy of (a) all genes and (b) low-expressing genes. Accuracy is defined as the deviation of pipeline-derived log ratios from the corresponding qPCR-based log ratios. It is encoded as color, with red representing the highest accuracy, or the lowest deviation from qPCR. "All Genes" refers to the 10,222 qPCR genes after filtering, and "Low-Expressing Genes" refers to the 2,044 qPCR genes after filtering.

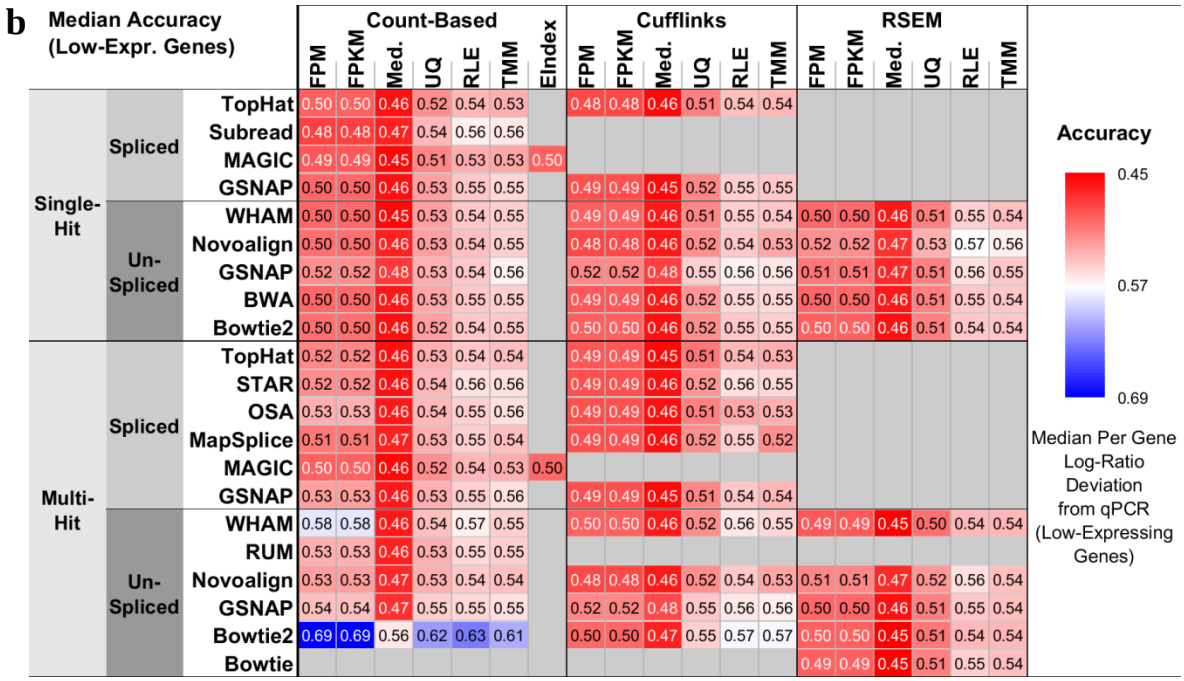


Figure 23 continued.

We observed the following results:

(1) Using all genes, the median log-ratio deviation between RNA-seq and qPCR ranged from 0.27 to 0.63 (**Figure 23**, panel a). Smaller deviation represents higher accuracy. Median normalization exhibited the lowest deviation, or the highest accuracy, compared with all other normalization methods. In addition, Cufflinks and RSEM performed similarly despite the choice of mapping algorithms. Moreover, for all mapping-quantification combinations, the [Bowtie2 multi-hit + count-based] pipelines showed the largest deviation. Furthermore, pipelines with multi-hit mapping and count-based quantification generally showed larger deviation than other pipelines. Among all pipeline factors, normalization was the largest statistically significant ($p<0.05$) source of variation (**Figure 24**, panel a).

(2) The median log-ratio deviation using low-expressing genes was larger than that using all genes, and it ranged from 0.45 to 0.69 (**Figure 23**, panel b). The trends of pipeline performance were similar to those using all genes, and normalization was also the largest statistically significant ($p<0.05$) source of variation (**Figure 24**, panel bError! Reference source not found.).

(3) In summary, median normalization with most mapping and quantification algorithms, besides [Bowtie2 multi-hit + count-based], was the best choice for quantifying genes with high accuracy, or low deviation from qPCR.

These results suggested that mapping, quantification, and normalization methods jointly impacted the accuracy of gene expression.

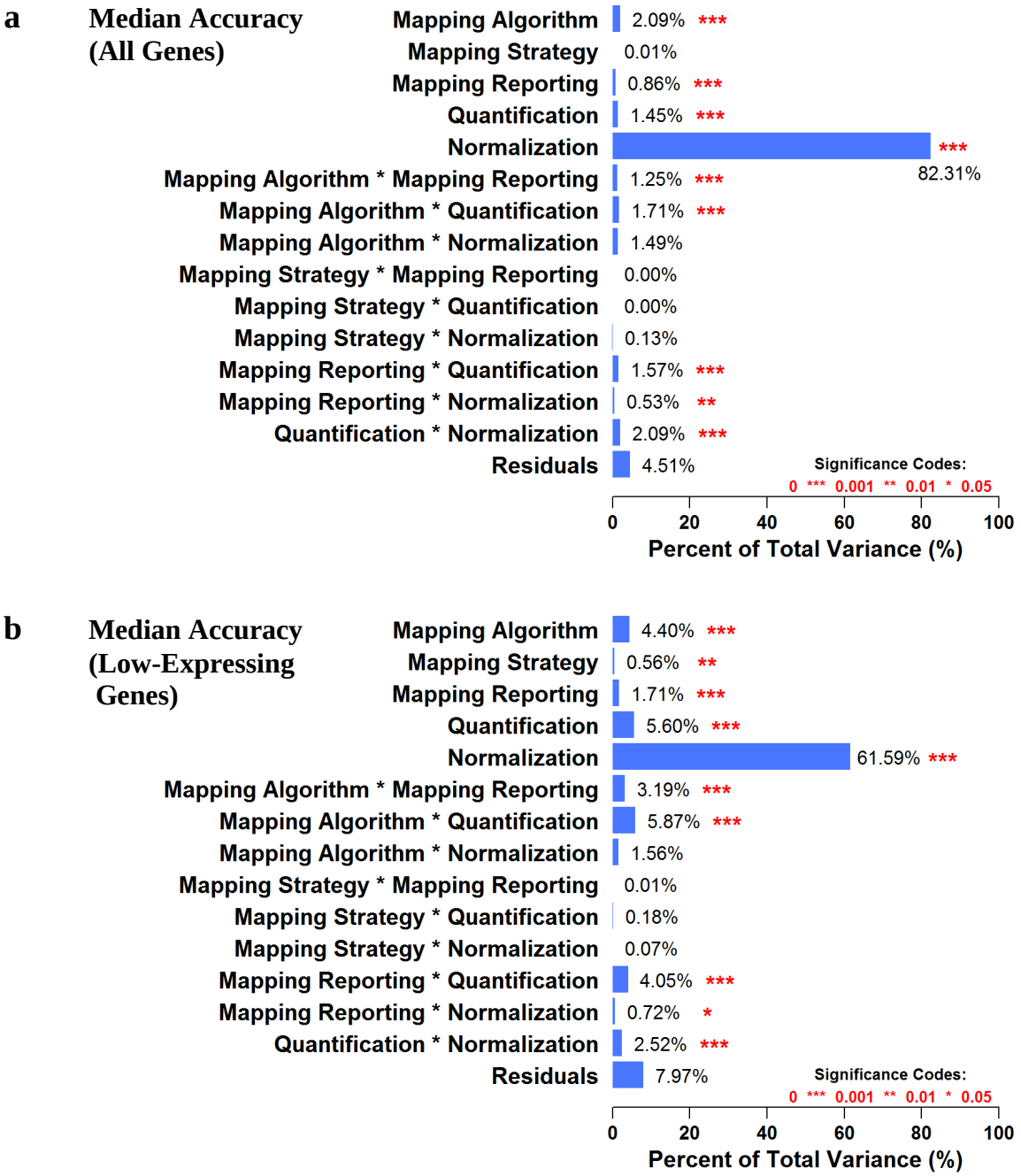


Figure 24: ANOVA for Median Accuracy of All and Low-Expressing Genes. Analysis of variance (ANOVA) decomposes the overall variance in the median accuracy of (a) all genes and (b) low-expressing genes into various factors considered, including RNA-seq pipeline components and associated two-way interactions. The statistical significance of the contribution of each component and interaction is denoted by red asterisks, with '***' indicates p-values are smaller than 0.001, '**' indicates p-values are smaller than 0.01, and '*' indicates p-values are smaller than 0.05. Among all factors, the normalization contributes the most to the overall variance.

Impact of mapping, quantification, and normalization on gene expression precision

We defined the precision metric as the coefficient of variation (CoV) of gene expression across replicate libraries, and visualized the median precision of all genes and low-expressing genes using heatmaps (**Figure 25**).

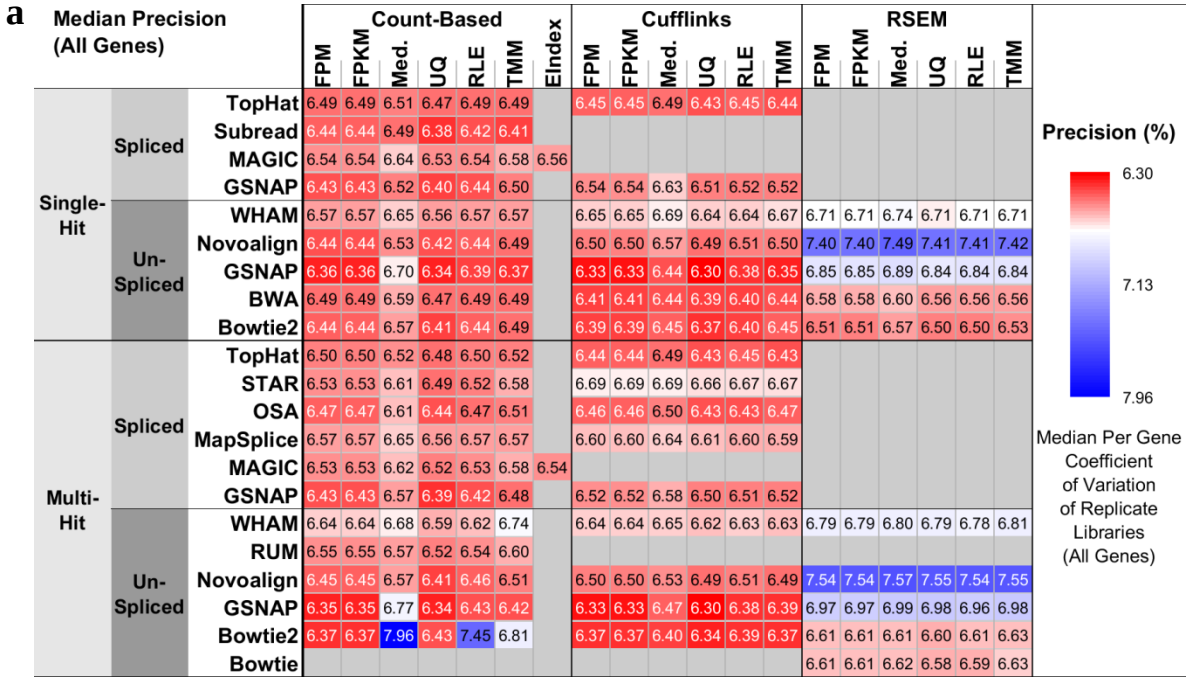


Figure 25: Median Precision of All and Low-Expressing Genes. The 278 RNA-seq pipelines applied to the SEQC-benchmark dataset differ in terms of the median precision of (a) all genes and (b) low-expressing genes. Precision is defined as the coefficient of variation over replicate libraries. It is encoded as color, with red representing the highest precision, or the lowest coefficient of variation.

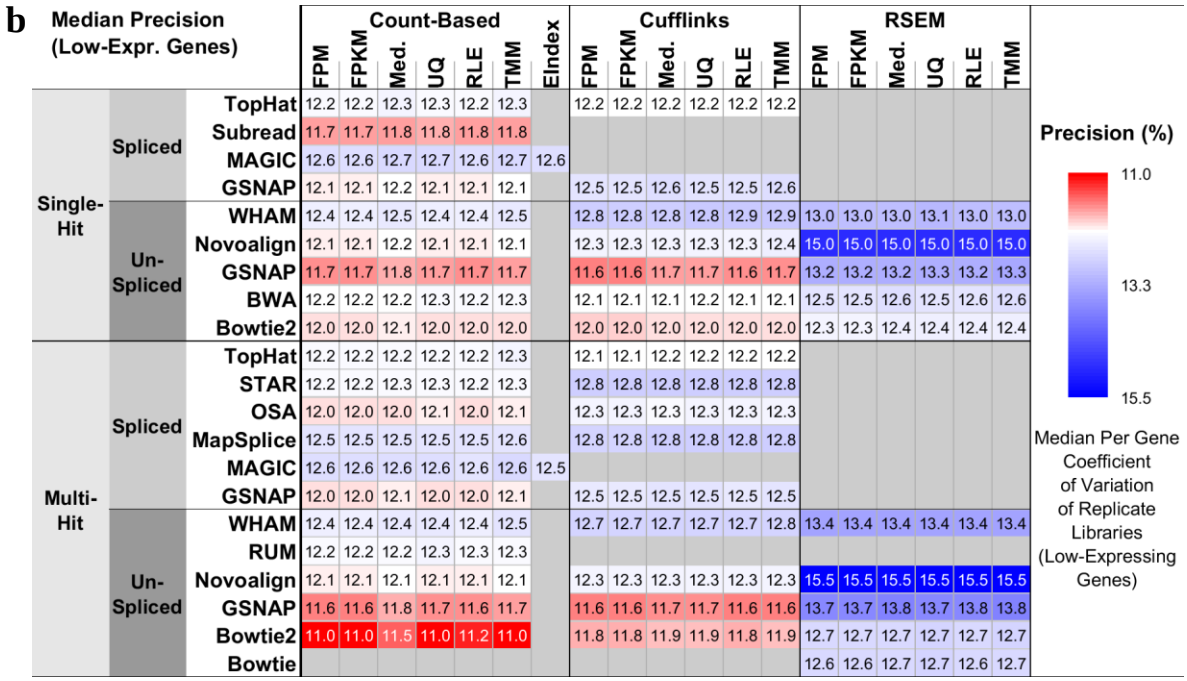


Figure 25 continued.

We observed the following results:

(1) Using all genes, the median CoV ranged from 6.30% to 7.96% (**Figure 25**, panel a). Smaller CoV represents higher precision. Pipelines with any of Novoalign, GSNAP un-spliced, or WHAM mapping, and RSEM quantification resulted in higher CoV, despite the choice of normalization methods. In addition, the [Bowtie2 multi-hit + count-based + med.] pipeline always led to the largest CoV. Moreover, for each mapping-normalization combination, pipelines with either count-based or Cufflinks quantification always reported higher precision than those with RSEM quantification, except the [Bowtie2 multi-hit + count-based + med.] pipeline. Quantification, mapping algorithm, and their interaction were the largest statistically significant ($p<0.05$) source of variation (**Figure 26**, panel a).

(2) The median CoV using low-expressing genes was larger than that using all genes, and it ranged from 11.0% to 15.5% (**Figure 25**, panel bError! Reference source not found.). The trends of pipeline performance were similar to those using all genes, except that the [Bowtie2 multi-hit + count-based] pipelines exhibited the highest precision among others. Again, quantification, mapping algorithm, and their interaction were the largest statistically significant ($p<0.05$) source of variation (**Figure 26**, panel b).

(3) In summary, pipelines with any of Bowtie2 multi-hit, GSNAP un-spliced, or Subread mapping and either count-based or Cufflinks quantification, besides the [Bowtie2 multi-hit + count-based + med.] pipeline, were the best choice for quantifying genes with high precision, or low CoV.

These results suggested that mapping, quantification, and normalization methods jointly affected the precision of gene expression.

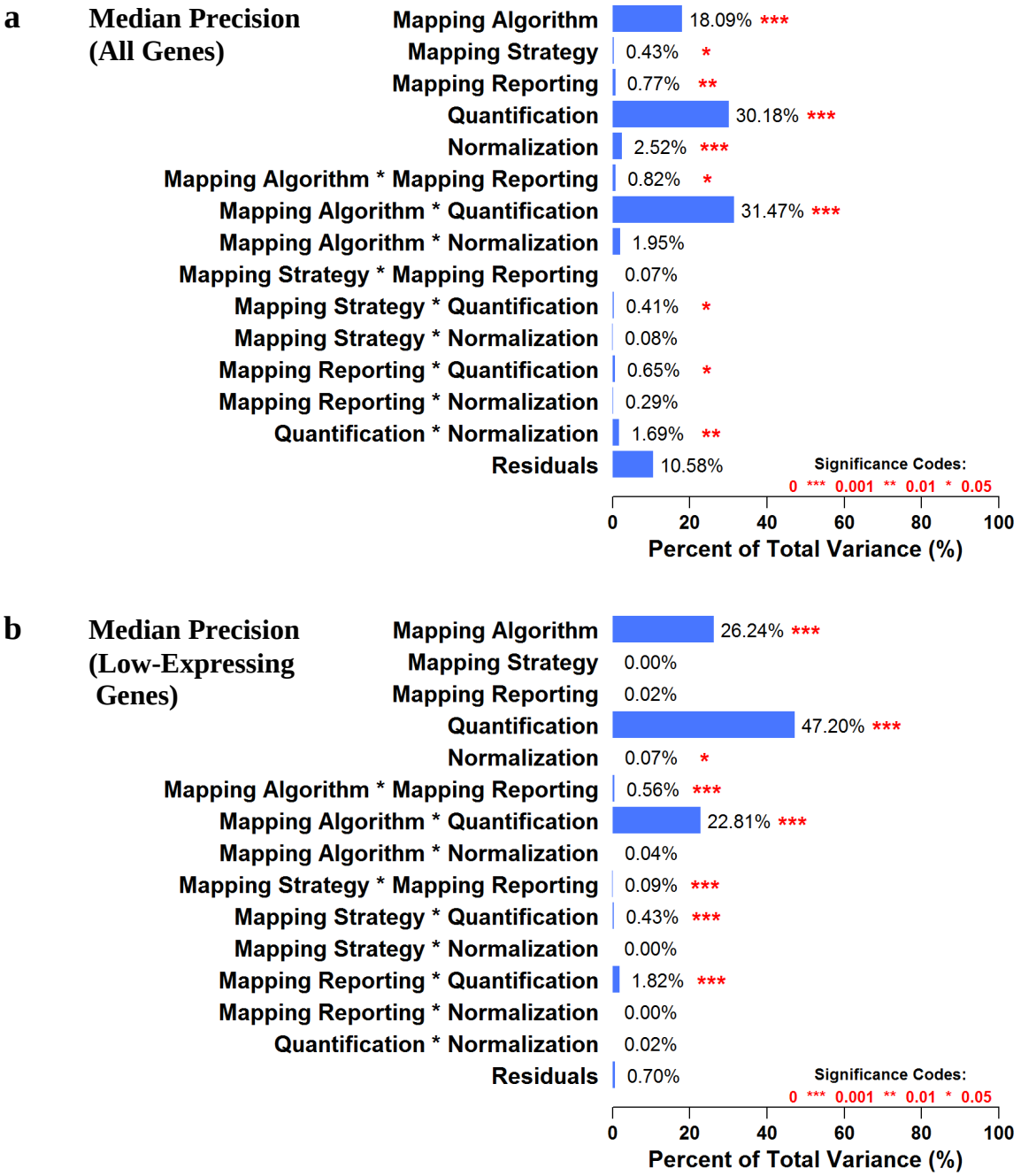


Figure 26: ANOVA for Median Precision of All and Low-Expressing Genes. Analysis of variance (ANOVA) decomposes the overall variance in the median precision of (a) all genes and (b) low-expressing genes into various factors considered, including RNA-seq pipeline components and associated two-way interactions. The statistical significance of the contribution of each component and interaction is denoted by red asterisks, with '***' indicates p-values are smaller than 0.001, '**' indicates p-values are smaller than 0.01, and '*' indicates p-values are smaller than 0.05. Among all components and interactions, the quantification, mapping algorithm, and mapping algorithm-quantification interaction contribute the most to the overall variance.

Impact of mapping, quantification, and normalization on gene expression reliability

We defined the reliability metric as the intraclass (i.e., intra-sample in the context of the SEQC-benchmark dataset) correlation (ICC) of gene expression, and visualized the median reliability of all genes and low-expressing genes using heatmaps (**Figure 27**).

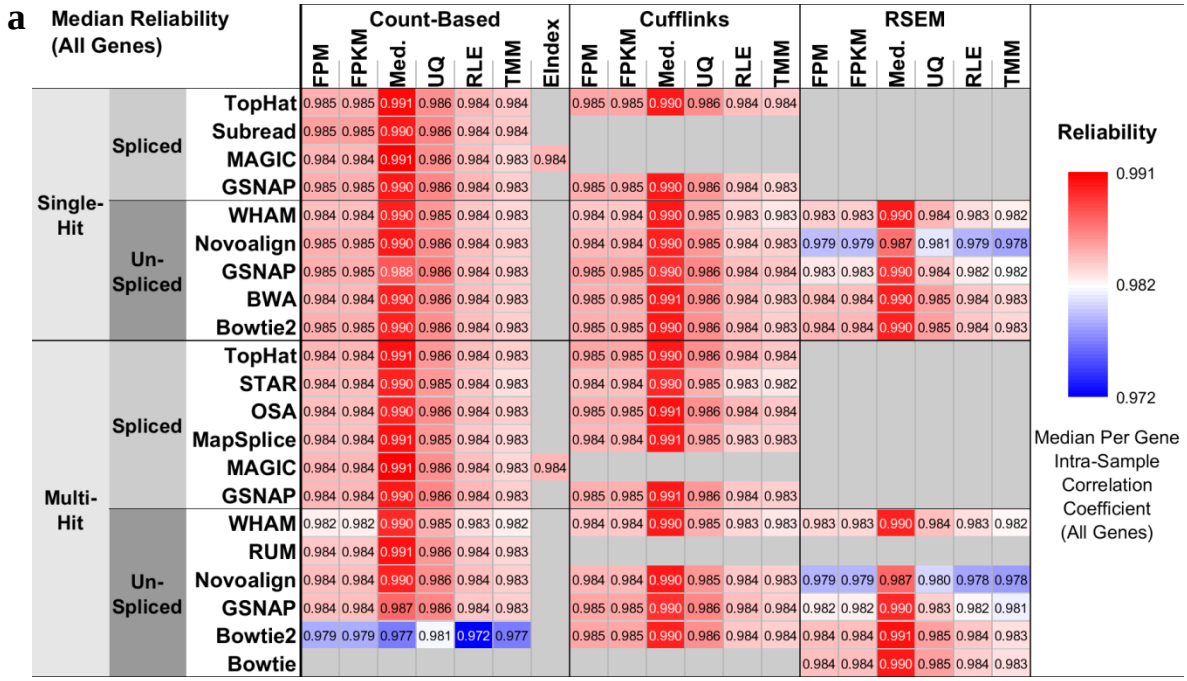


Figure 27: Median Reliability of All and Low-Expressing Genes. The 278 RNA-seq pipelines applied to the SEQC-benchmark dataset differ in terms of the median reliability of (a) all genes and (b) low-expressing genes. Reliability is defined as the intraclass (or intra-sample in our case) correlation that quantifies how similar replicate libraries of a sample are to one another using analysis of variance techniques. It is encoded as color, with red representing the highest reliability, or the highest intraclass correlation.

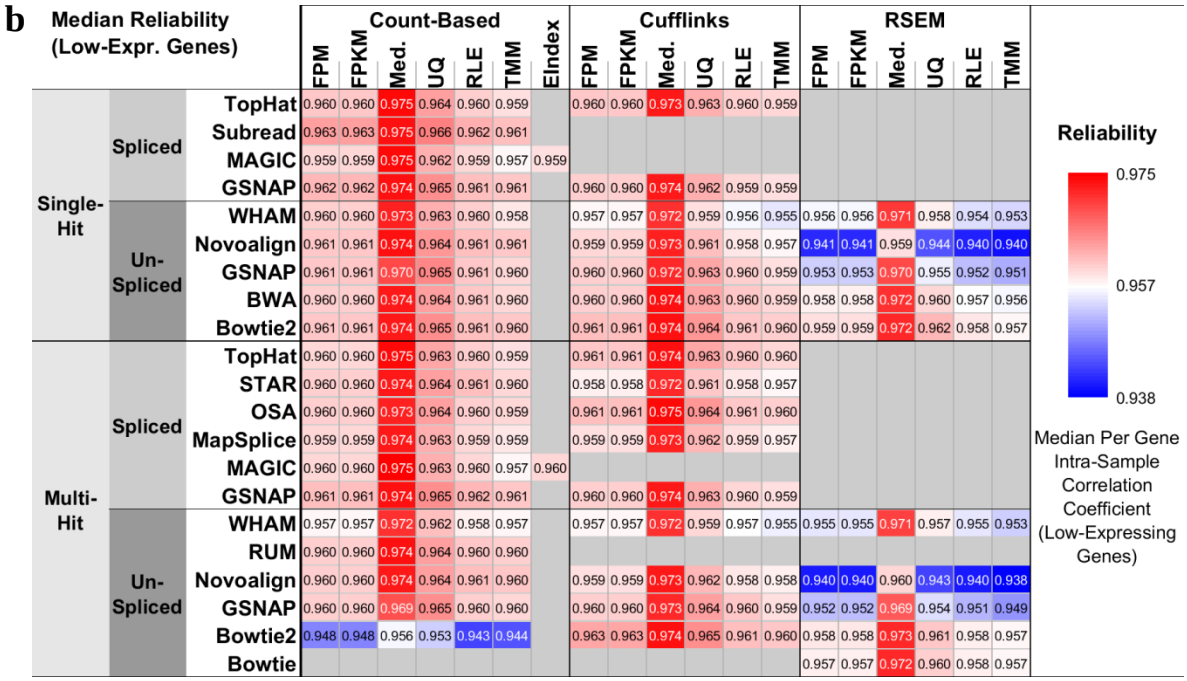


Figure 27 continued.

We observed the following results:

(1) Using all genes, the median ICC ranged from 0.972 to 0.991 (**Figure 27**, panel a). Larger ICC represents higher reliability. Median normalization exhibited the highest ICC, or the highest reliability, compared with all other normalization methods. In addition, pipelines with Novoalign mapping and RSEM quantification resulted in lower ICC despite the choice of normalization methods. Moreover, the [Bowtie2 multi-hit + count-based] pipelines showed the lowest ICC. Furthermore, for each mapping-normalization combination, pipelines with either count-based or Cufflinks quantification always reported higher ICC than those with RSEM quantification, except the [Bowtie2 multi-hit + count-based] pipelines mentioned previously. Normalization was the largest statistically significant

($p<0.05$) source of variation, followed by two-way [mapping algorithm*quantification] interaction (**Figure 28**, panel a).

(2) The median ICC using low-expressing genes was smaller than that using all genes, and it ranged from 0.938 to 0.975 (**Figure 27**, panel bError! Reference source not found.). The trends of pipeline performance were similar to those using all genes, except [Novoalign + RSEM] pipelines and [Bowtie2 multi-hit + count-based] pipelines. Normalization, two-way [mapping algorithm*quantification] interaction, quantification, and mapping algorithm were the largest statistically significant ($p<0.05$) source of variation (**Figure 28**, panel b).

(3) In summary, median normalization along with most mapping and quantification algorithms, besides the [Bowtie2 multi-hit + count-based] and [Novoalign + RSEM] pipelines, was the best choice for quantifying genes with high reliability, or high ICC.

These results suggested that mapping, quantification, and normalization methods jointly affected the reliability of gene expression.

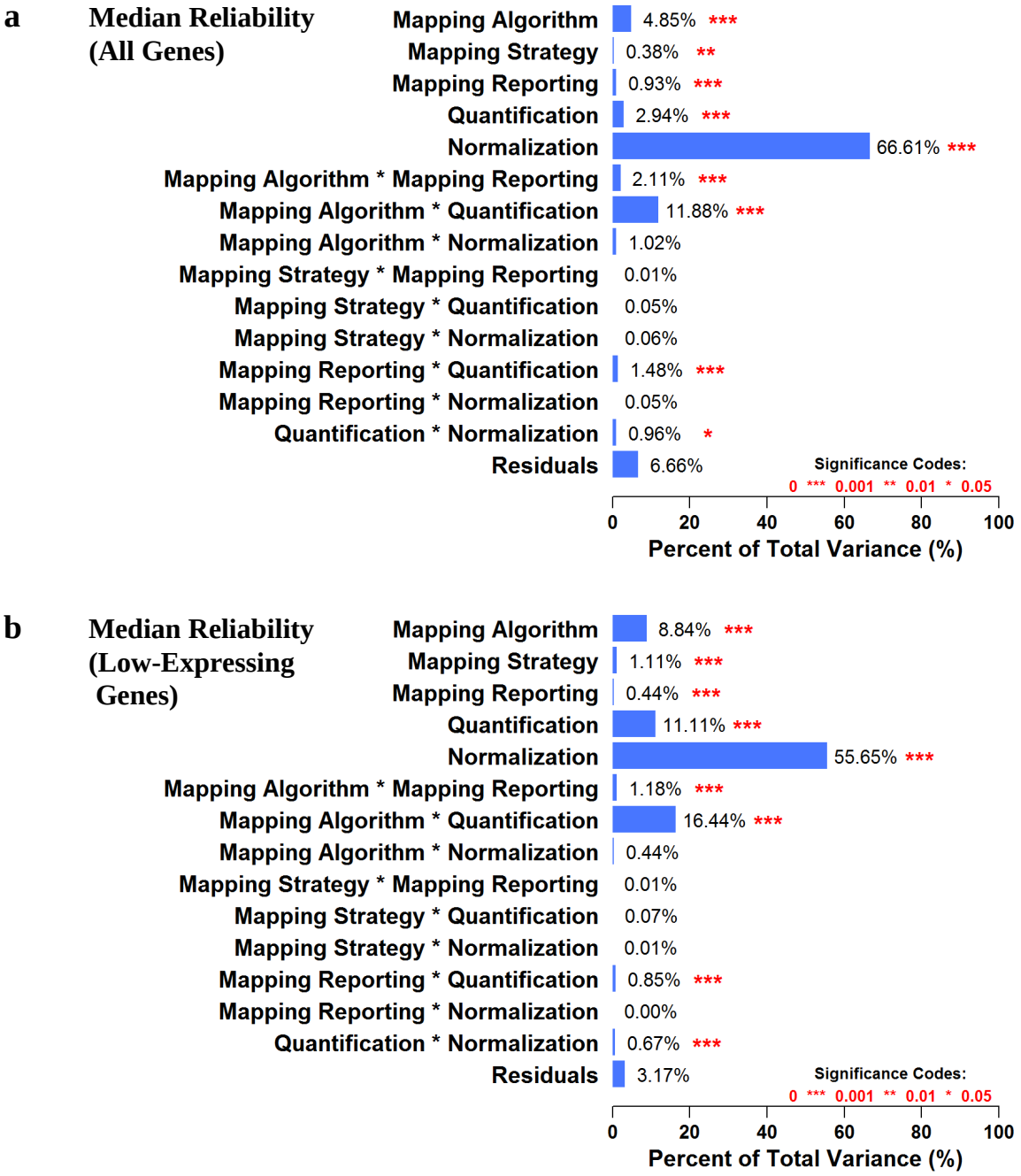


Figure 28: ANOVA for Median Reliability of All and Low-Expressing Genes. Analysis of variance (ANOVA) decomposes the overall variance in the median reliability of (a) all genes and (b) low-expressing genes into various factors considered, including RNA-seq pipeline components and associated two-way interactions. The statistical significance of the contribution of each component and interaction is denoted by red asterisks, with '***' indicates p-values are smaller than 0.001, '**' indicates p-values are

smaller than 0.01, and '*' indicates p-values are smaller than 0.05. Among all factors, the normalization contributes the most to the overall variance.

Impact of mapping, quantification, and normalization on gene expression reproducibility

We defined the reproducibility metric as the Spearman correlation between two replicate libraries of the same sample, and visualized the median reproducibility of all genes and low-expressing genes using heatmaps (**Figure 29**).

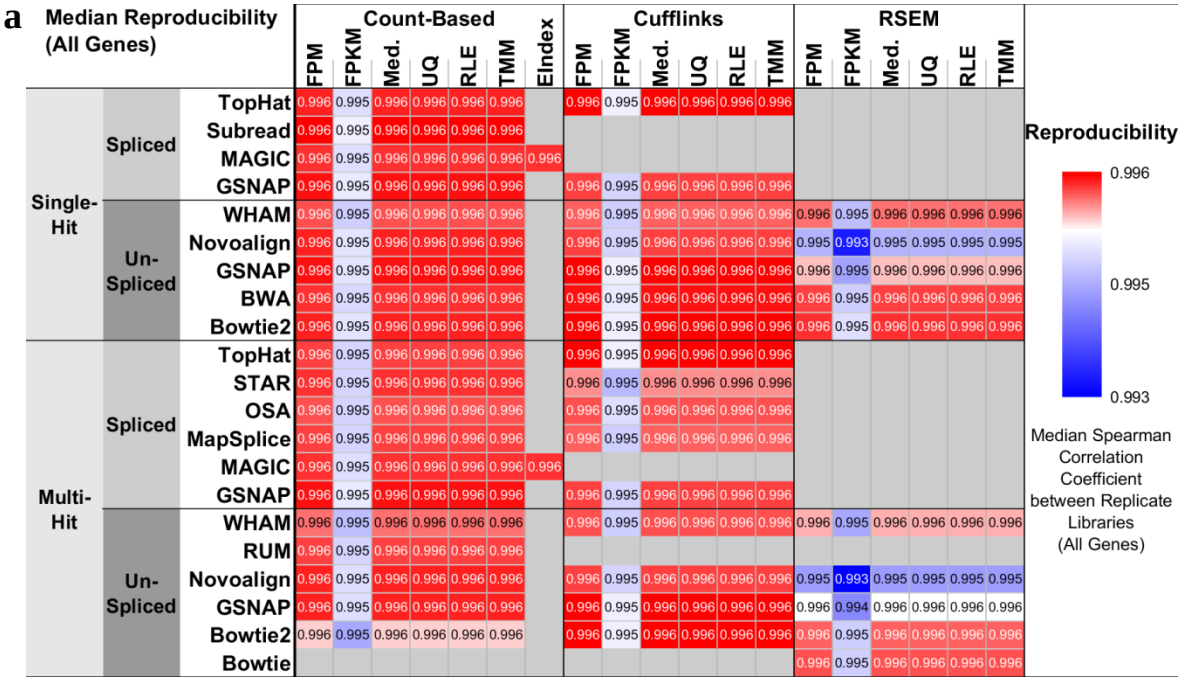


Figure 29: Median Reproducibility of All and Low-Expressing Genes. The 278 RNA-seq pipelines applied to the SEQC-benchmark dataset differ in terms of the median reproducibility of (a) all genes and (b) low-expressing genes. Reproducibility is defined as the Spearman correlation between two replicate libraries of the same sample. It is encoded as color, with red representing the highest reproducibility, or the highest Spearman correlation.

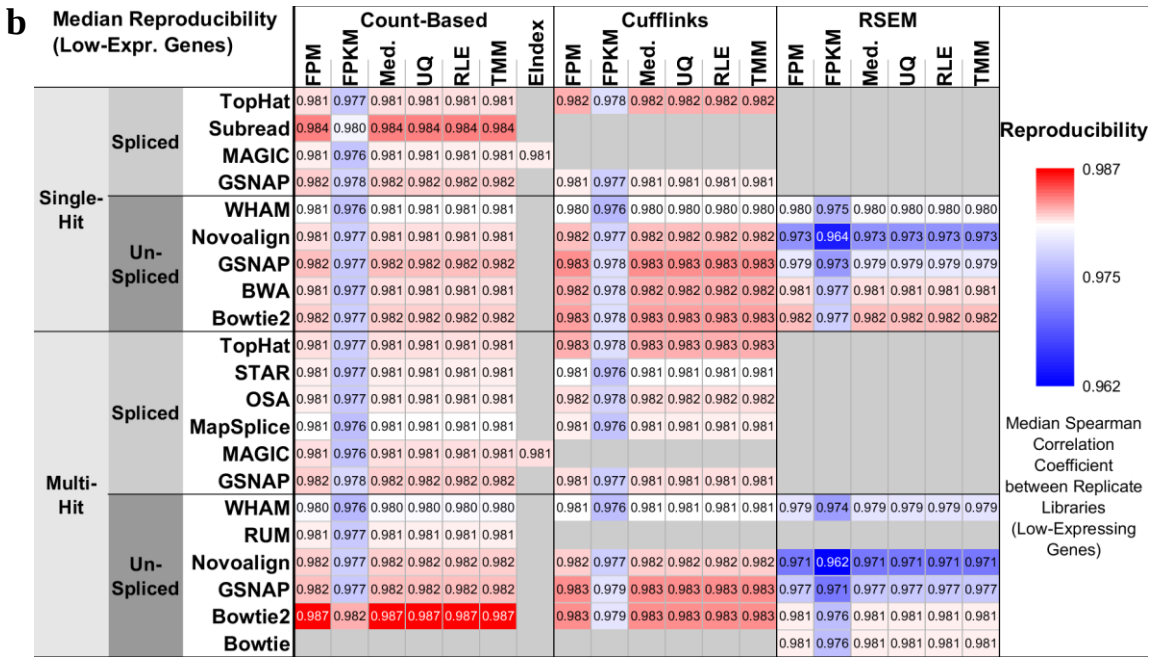


Figure 29 continued.

We observed the following results:

(1) Using all genes, the median Spearman correlation ranged from 0.993 to 0.996 (**Figure 29**, panel a). Larger Spearman correlation represents higher reproducibility. Since the Spearman correlation is based on the rank values of gene expression, normalization methods scaling all expression with the same factor (i.e., all but the FPKM method) will lead to the same reproducibility values. The FPKM normalization method always exhibited the lower Spearman correlation, or the lower reproducibility, than the other normalization methods. In addition, pipelines with either Novoalign or GSNAP un-spliced mapping and RSEM quantification resulted in lower Spearman correlation despite the choice of normalization methods.

Normalization, two-way [mapping algorithm*quantification] interaction, quantification, and mapping algorithm were the largest statistically significant ($p<0.05$) source of variation (**Figure 30**, panel a).

(2) The median Spearman correlation using low-expressing genes was smaller than that using all genes, and it ranged from 0.962 to 0.987 (**Figure 29**, panel bError! Reference source not found.). The trends of pipeline performance were similar to those using all genes. Normalization, quantification, two-way [mapping algorithm*quantification] interaction, and mapping algorithm were the largest statistically significant ($p<0.05$) source of variation (**Figure 30**, panel b).

(3) In summary, all but the FPKM normalization method with either Subread mapping and count-based quantification or GSNAP un-spliced mapping and Cufflinks quantification were the best choice for quantifying genes with high reproducibility, or high Spearman correlation.

These results suggested that mapping, quantification, and normalization methods jointly affected the reproducibility of gene expression.

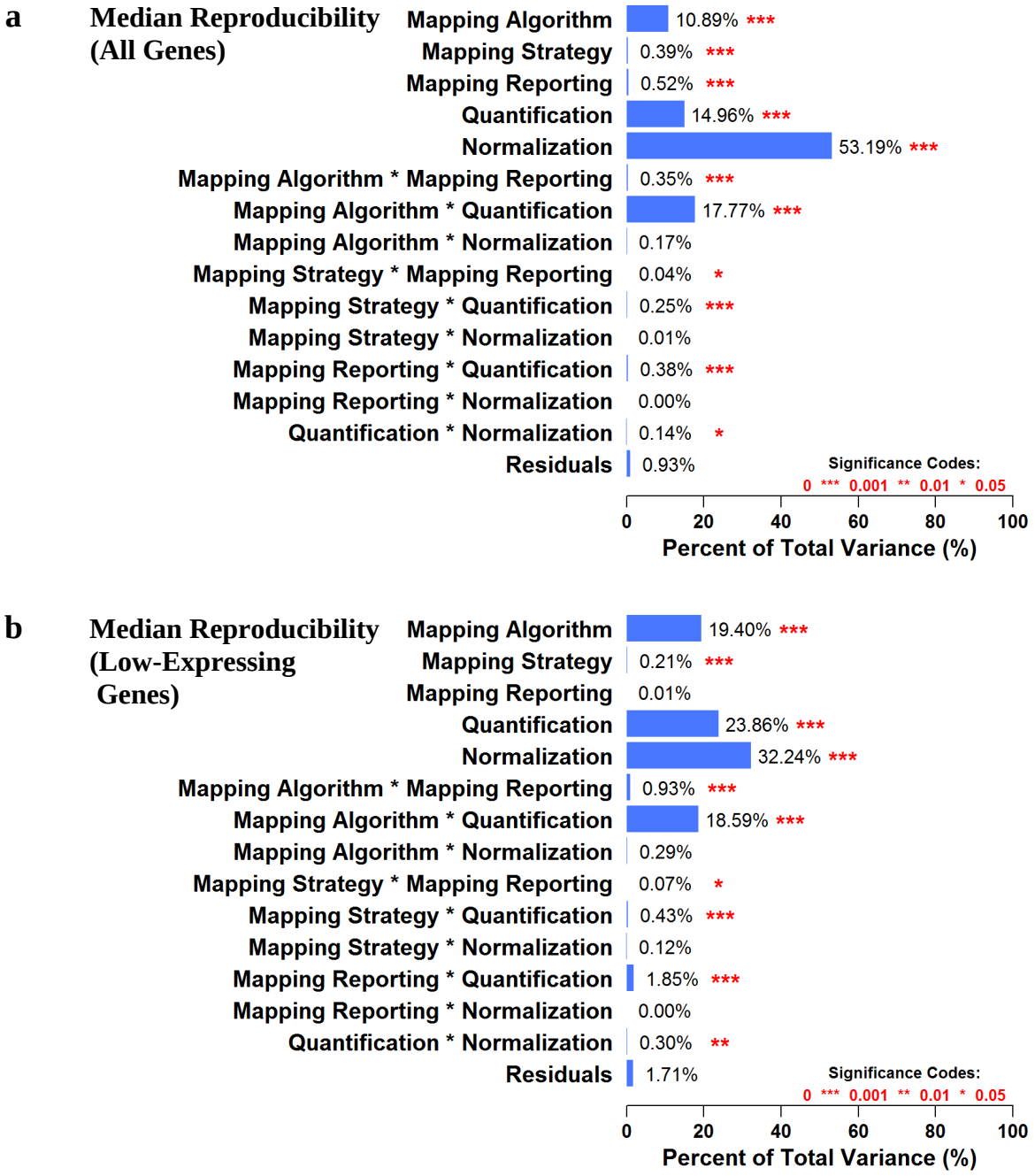


Figure 30: ANOVA for Median Reproducibility of All and Low-Expressing Genes. Analysis of variance (ANOVA) decomposes the overall variance in the median reproducibility of (a) all genes and (b) low-expressing genes into various factors considered, including RNA-seq pipeline components and associated two-way interactions. The statistical significance of the contribution of each component and

interaction is denoted by red asterisks, with '***' indicates p-values are smaller than 0.001, '**' indicates p-values are smaller than 0.01, and '*' indicates p-values are smaller than 0.05. Among all components and interactions, the normalization contributes the most to the overall variance.

Investigating relationship between alignment profiles and benchmark metrics

The performance of benchmark metrics depended on characteristics of mapping results. We used the M-estimation with Huber weighting to fit linear models that capture the relationship between the benchmark metrics and alignment profiles (see the Appendix B section "Regression Analysis" for details). The accuracy metric correlated with the number of mismatches per mapped read, and the precision, reliability, and reproducibility metrics correlated with the number of mapped fragments (**Figure 31**). Less mismatches per read and more mapped fragments led to more accurate, precise, reliable, and reproducibility gene expression.

The Case Study 4 investigation using the SEQC-benchmark dataset demonstrated that gene expression estimation is significantly impacted by the joint effect of multiple RNA-seq pipeline components (**Figure 23** to **Figure 30**Error! Reference source not found.).

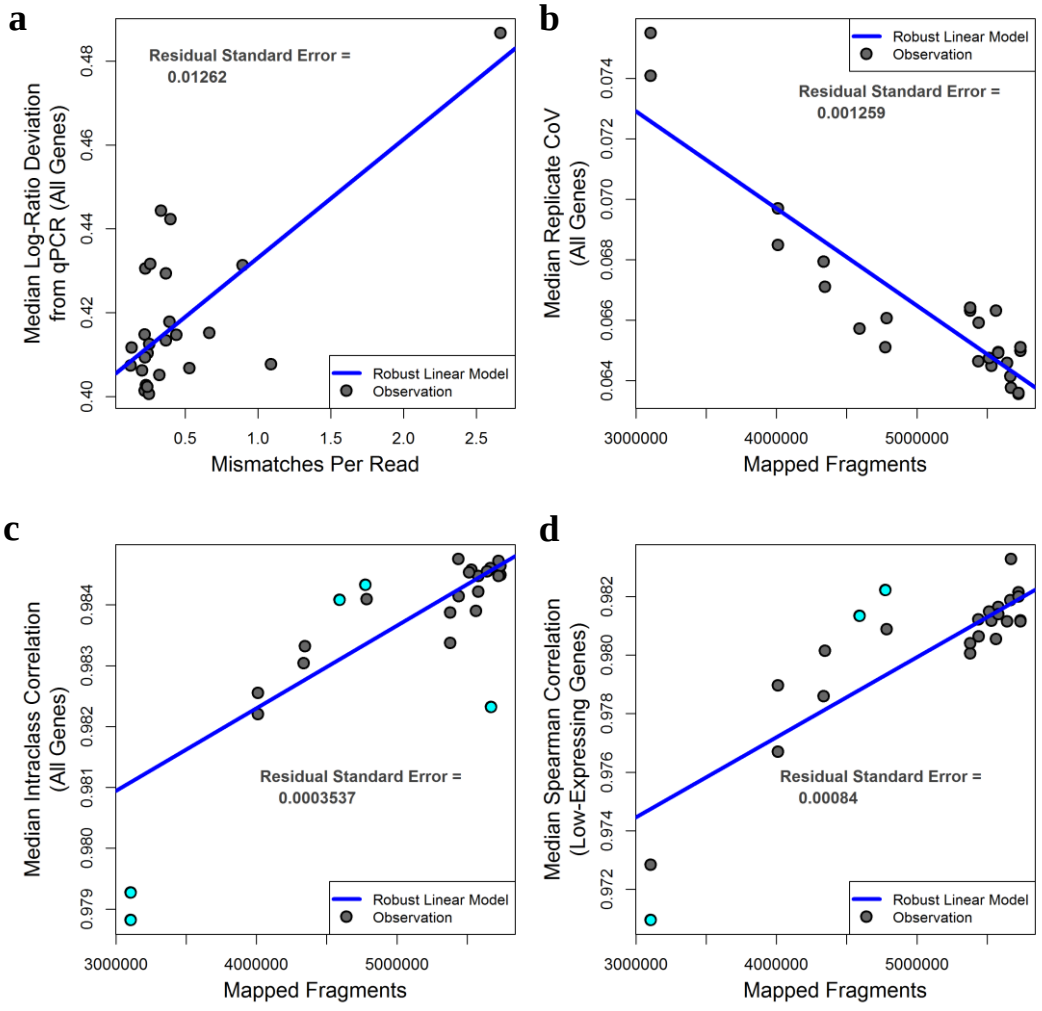


Figure 31: Relationship between Alignment Profiles and Benchmark Metrics. Benchmark metric performance correlates with alignment profiles. For each panel, the x-axis represents an alignment profile, and the y-axis corresponds to a benchmark metric. Each gray point represents a sequence mapping pipeline, and the blue line depicts the robust linear model using M-estimation with Huber weighting. Points in cyan have Huber weights less than 0.5 (i.e., potential outlying points). (a) The median deviation of all genes positively correlates with the number of mismatches per read. More mismatches per read results in higher deviation, or lower accuracy; (b) the median coefficient of variation (CoV) of all genes negatively correlates with the number of mapped fragments. More mapped fragments results in lower CoV, or higher precision; (c) the median

intraclass correlation of all genes positively correlates with the number of mapped fragments. More mapped fragments results in higher intraclass correlation, or higher reliability; and finally (d) the median Spearman correlation of low-expressing genes positively correlates with the number of mapped fragments. More mapped fragments results in higher Spearman correlation, or higher reproducibility.

2.4.4.5 Summary of Case Study

We performed a systematic investigation of the 278 representative RNA-seq pipelines that resulted in multiple resources for RNA-seq users. In Case Study 4, after the SEQC conducted a broad investigation of RNA-seq technology [148], we developed a set of metrics to characterize RNA-seq pipelines using the SEQC-benchmark dataset. These metrics included the deviation of gene expression from qPCR data that quantifies accuracy; the CoV of gene expression across replicate libraries that quantifies precision; the ICC of gene expression grouped by samples that quantifies reliability; and the Spearman correlation between replicate libraries that quantifies reproducibility. We observed that RNA-seq pipeline components jointly affected gene expression estimation (**Figure 23** to **Figure 30**). These joint effects had not previously been reported in studies that investigated individual RNA-seq pipeline components.

We summarize and compare the results of our study to previous studies focusing on individual pipeline components. For example, previous studies observed that RUM, GSNAP spliced, STAR, and MapSplice mapping led to more accurate base-level alignment and splice junction detection [128, 150]. In addition, BWA, Bowtie, and Bowtie2 mapping were reported to be robust to sequencing errors and indels [149]. We similarly observed considerable differences in alignment profiles among mapping algorithms, and such the differences led to variations in the benchmark metrics (**Figure 31**). For example, Bowtie2 multi-hit mapping aligned many more reads, a higher

percentage of which were sub-optimal mapping variants (i.e., secondary mappings, mismatches, insertions, deletions, and splicing), than WHAM single-hit mapping (**Figure 32**, panels a – c). Consequently, pipelines with Bowtie2 multi-hit mapping resulted in larger deviation from the qPCR reference, or lower accuracy, than those with WHAM single-hit mapping. However, such the observation applied to only count-based quantification but not Cufflinks or RSEM (**Figure 23**). In addition to observations corresponding to previous literature, we also observed a joint effect between mapping and quantification components.

Variations in mapping performance propagated to the quantification stage. Quantification strategy for multi-hit mappers may explain the variation in gene expression accuracy. For example, Cufflinks and RSEM use Poisson distribution-based models and assign probabilities to each mapping while HTSeq simply counts total mapped reads regardless of quality. Thus, Cufflinks and RSEM are better able to handle multi-hit information, resulting in smaller deviation from the qPCR reference (**Figure 23** and **Figure 32**, panel d).

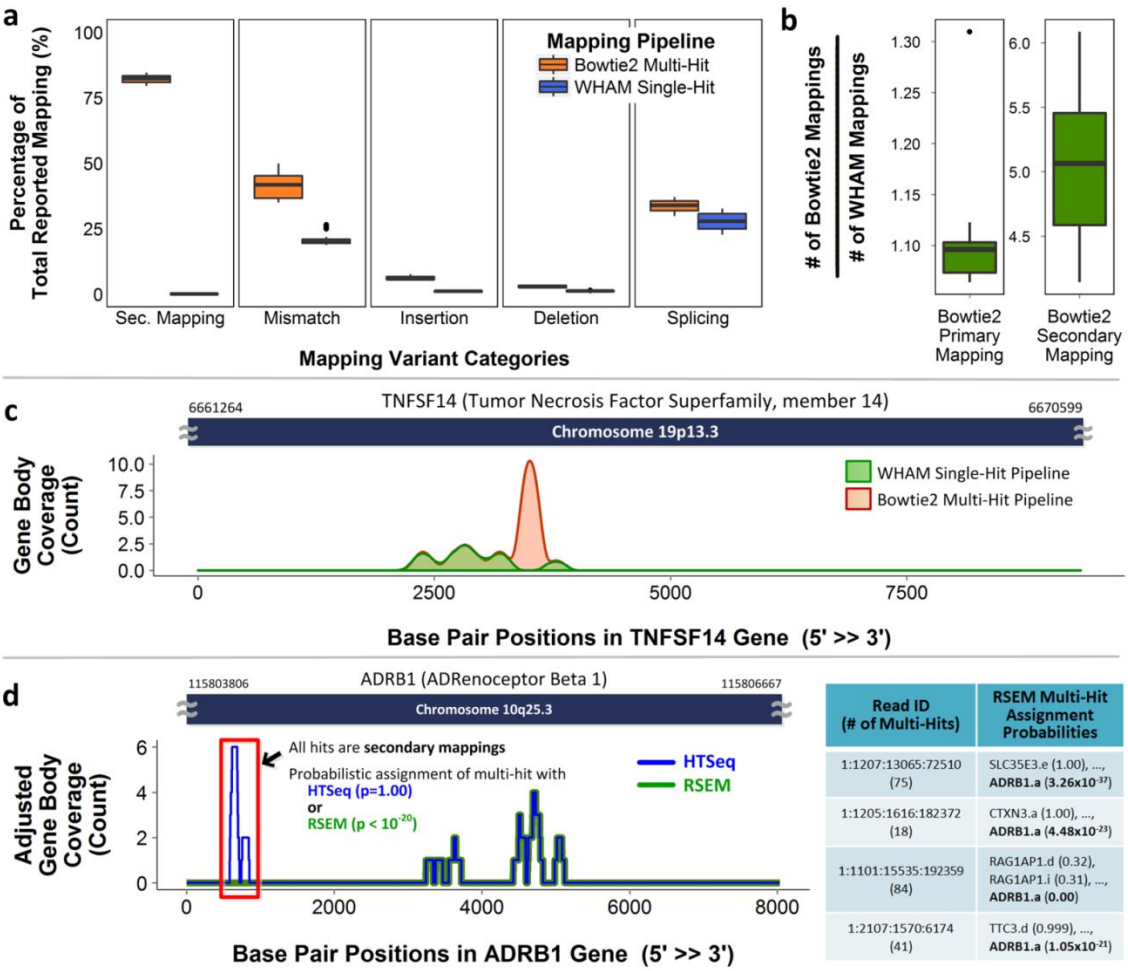


Figure 32: The Impact of Pipeline Choices on Mapping and Quantification Outcome. Bowtie2 multi-hit and WHAM single-hit pipelines differ significantly in terms of the percentage of total reported mappings for each mapping variant (i.e., secondary mapping, mapping with mismatches, insertion, deletion, and splicing). Each box demonstrates the distribution of percentages calculated from multiple sample replicates in the SEQC-benchmark dataset. The Bowtie2 multi-hit pipeline reports a higher percentage of mapping variants than the WHAM single-hit pipeline. (b) The WHAM single-hit pipeline reports only primary mappings. The box plot shows the distribution of the ratios of total primary or secondary mappings of the Bowtie2 multi-hit pipeline to total primary mappings of the WHAM single-hit pipeline using multiple sample replicates in the SEQC-benchmark dataset. The Bowtie2 multi-hit pipeline reports slightly more primary mappings than the WHAM single-hit pipeline, and it reports approximately five times more secondary mappings than the WHAM single-hit pipeline. These additional secondary mappings are informative for some expression quantification algorithms. (c) Gene body coverage differs between the Bowtie2 multi-hit and WHAM single-hit pipelines. The former pipeline reports additional secondary mappings between position 3200 and 3800. (d) Gene body coverage adjusted by multi-hit assignment probabilities derived from the HTSeq or RSEM quantification pipeline.

2.5 Summary and Key Innovations

In this chapter, I have addressed the first specific aim of this dissertation by designing several experiments (i.e., case studies) and evaluation metrics that can facilitate quality control of gene expression estimation. At the beginning of this chapter, I echoed Section 1.5 with the more specific introduction of RNA-seq feature extraction pipelines relevant to the four case studies described later in this chapter. Then, I detailed many evaluation metrics I designed for assessing the performance of these feature expression pipelines. Lastly, I elaborated on the experimental design, datasets, results, discussion, and conclusion of the four case studies.

The first three case studies focused on investigating the effect of each individual pipeline component, including the genome annotation for sequence mapping, the quantification pipeline, and the normalization method. **Figure 33** summarizes recommended RNA-seq pipeline components based on practical objectives. In contrast, the last case study (i.e., Case Study 4) emphasized on the joint effect of pipeline components on gene expression quality. The experimental design of this case study was much more complicated than the other three, and **Figure 34** summarizes good-performing RNA-seq pipelines for various RNA-seq applications.

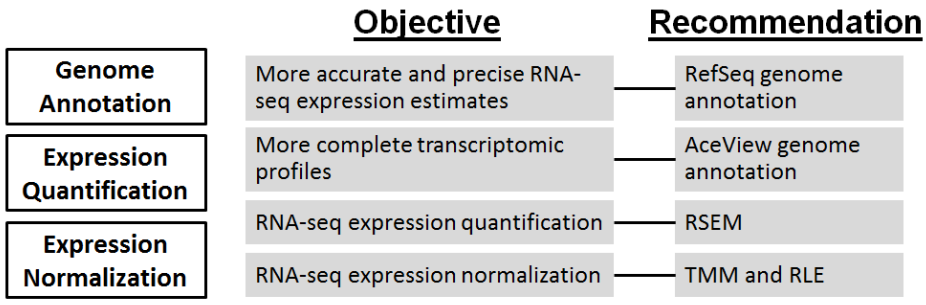


Figure 33: Summary of Component-wise Investigation and Recommendation.

Good-Performing RNA-seq Pipelines for Various Applications

RNA-seq Application	Metric	RNA-seq Pipelines
Accurate estimation of relative gene expression with benefit for differentially expressed gene detection	**Accuracy** (Deviation from qPCR)	• Bowtie + RSEM + Median • Bowtie2 Single-Hit + [Count-Based/Cufflinks/RSEM] + Median • Bowtie2 Multi-Hit + RSEM + Median • BWA + [Count-Based/RSEM] + Median • GSNAP Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • GSNAP Un-spliced Multi-Hit + RSEM + Median • MAGIC [Single-/Multi-Hit] + Count-Based + Median • OSA + [Count-Based/Cufflinks] + Median • STAR + Count-Based + Median • TopHat [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • WHAM Single-Hit + Count-Based + Median • WHAM Multi-Hit + RSEM + Median
Small variation in gene expression across all replicate libraries for a single sample	**Precision** (Coefficient of variation across replicate libraries)	• Bowtie2 Multi-Hit + Count-Based + [FPM/FPKM/Upper Quartile] • Bowtie2 Multi-Hit + Cufflinks + RLE • GSNAP Un-Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + [FPM/FPKM/Upper Quartile/RLE] • GSNAP Un-Spliced [Single-/Multi-Hit] + Cufflinks + TMM
Small within-sample variation in gene expression across all replicate libraries compared with between-sample variation	**Reliability** (Intraclass [intra-sample] correlation for grouped data)	• Bowtie2 [Single-/Multi-Hit] + [Count-Based/Cufflinks/RSEM] + Median • BWA + [Count-Based/Cufflinks/RSEM] + Median • GSNAP Spliced [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • MAGIC [Single-/Multi-Hit] + Count-Based + Median • MapSplice + [Count-Based/Cufflinks] + Median • Novoalign [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median • OSA + Cufflinks + Median • RUM + Count-Based + Median • Subread + Count-Based + Median • TopHat [Single-/Multi-Hit] + [Count-Based/Cufflinks] + Median
Small variation in the rank of gene expression between replicate libraries for a single sample	**Reproducibility** (Spearman correlation between replicate libraries)	• Bowtie2 [Single-/Multi-Hit] + Cufflinks + [FPM/Median/Upper Quartile/RLE/TMM] • GSNAP Un-Spliced [Single-/Multi-Hit] + Cufflinks + [FPM/Median/Upper Quartile/RLE/TMM] • TopHat [Single-/Multi-Hit] + Cufflinks + [FPM/Median/Upper Quartile/RLE/TMM]

Figure 34: Summary of Pipeline-wise Investigation and Recommendation. The resources provided by this study (i.e., the 278 RNA-seq pipelines, the benchmark metrics, and the SEQC-benchmark datasets) can serve as guidelines for biological and clinical researchers as well as for bioinformaticians and biotechnologists. Depending on the gene expression application, the accuracy, precision, reliability, and reproducibility metrics may be used to choose a pipeline. We have associated each metric with an RNA-seq application and listed the top-performing pipelines for each metric. The red-highlighted component in each listed RNA-seq pipeline indicates components that occur frequently among the top-performing pipelines for each metric.

The key innovations of the work in this chapter are listed as follows:

- I designed a comprehensive list of evaluation metrics that capture the performance of RNA-seq expression analysis pipeline.
- I conducted the first investigation on genome annotation and proposed a novel, informative annotation complexity measure.
- I performed quantification pipeline investigation (among the first batch) and identified key factors for achieving accurate expression estimates.

- I accomplished the simulation-based investigation on expression normalization (among the first batch).
- I performed the largest investigation of RNA-seq expression analysis pipeline so far using well-designed benchmark datasets provided by FDA.

CHAPTER 3

KNOWLEDGE DISCOVERY FOR PRECISION MEDICINE

3.1 Introduction

As discussed in Chapter 1, this dissertation aims to promote precision medicine by addressing major challenges in the three directions—quality control, knowledge discovery, and integrative analysis. Challenges associated with quality control of gene expression estimation have been addressed in Chapter 2. With good gene expression quality, more reliable knowledge can be discovered. The second specific aim of this dissertation was to discover impactful biomarkers that can facilitate subgroup assignment using NGS data. This aim can be addressed from two perspectives—model construction, which establishes models that can ultimately classify patients into disease subgroups, and biomarker identification and interpretation, which reports statistically significant or predictive biomarkers that may be applicable in clinical settings.

In this chapter, I first introduce methods for biomarker identification and predictive modeling specifically tailored to NGS data. Next, I use two case studies to demonstrate these methods and discuss findings from the two case studies. The background, experimental design, datasets, and results of the first case study are discussed based on its original publication [157], and those of the second case study are in preparation for submission to *Nature Methods*.

3.2 Biomarker Identification and Predictive Modeling for NGS Data

After feature extraction, raw RNA-seq data can be represented as gene expression tables with typically tens of thousands of gene expression for tens or hundreds of samples. Knowledge discovery for RNA-seq gene expression tables can be two-fold—identifying statistically significant biomarkers using statistical approaches or constructing predictive models using classification (a.k.a. supervised learning) techniques, both require predefined labels (i.e., known clinical outcome) for each sample in the study. In this section, I cover both perspectives. For the statistical modeling part, I first introduce how DEG detection works, followed by methods for protein DNA-binding site identification using chromatin immunoprecipitation sequencing (ChIP-seq) data. For the supervised learning part, I introduce a nested cross-validation technique I applied to estimate the prediction performance of clinical endpoints of interest.

3.2.1 Differentially Expressed Gene Detection

One most popular application of RNA-seq is to detect DEGs between two or more groups of samples. RNA-seq expression estimates from the two or more groups are first fit to a statistical distribution, followed by a statistical hypothesis test that determines whether the distributions between (among) the two or more groups are statistically significantly different for a targeted gene. Soneson *et al.* and Rapaport *et al.* have comprehensively conducted quantitative evaluation for DEG detection methods [158, 159]. Thus, this section will mainly focus on qualitative categorization.

DEG detection methods can be nonparametric or parametric. Nonparametric methods such as SAMseq [160] and NOISeq [161] use resampling and counting techniques to avoid making assumptions about the underlying distribution of RNA-seq

expression estimates. Data permutation is a common technique for estimating false discovery rates for nonparametric methods.

Parametric methods use Poisson-based models to fit RNA-seq read count data [162]. The Poisson distribution has the variance equals to the mean. However, overdispersion (i.e., when the variance is significantly greater than the mean) often occurs in RNA-seq data. Therefore, the negative binomial distribution, which is a two-parameter extension of the Poisson distribution, introduces an additional parameter to capture the high variability [98]. Selecting an appropriate statistical model is the key for a parametric DEG detection method. For example, DEGseq [163] applies the Poisson distribution to model RNA-seq read count data; edgeR [65], baySeq [164], and DESeq [98] use the negative binomial model to capture the overdispersion; Myrna [165] models the data as either the Gaussian distribution or the Poisson distribution; and Cuffdiff2 [166] uses the beta negative binomial model to capture both overdispersion and uncertainty in the fragment count of a transcript. After constructing the statistical model, most parametric methods assess the significance level of each gene by using either p-values computed from the likelihood ratio test or the Fisher's exact test, or posterior probabilities estimated from the empirical Bayes method, while Cuffdiff2 assumes that RNA-seq data is normally distributed after a particular transformation and uses the t-test to determine the statistical significance of each DEG.

3.2.2 Protein DNA-Binding Site Identification

The bioinformatics pipeline for protein DNA-binding site identification using ChIP-seq data is composed of two steps—sequence mapping, which is very similar to that for RNA-seq data, and peak calling, which determine statistically significant peaks in

a dataset. Sequence reads generated from ChIP-seq mostly originate from DNA sequences around targeted protein DNA-binding regions. After mapping reads to the reference genome and identifying uniquely mapped reads, genomic loci that accumulate a large number of reads (i.e., peaks) indicate putative protein DNA-binding regions. Peak-calling tools distinguish true peaks from background noise by (1) generating a signal profile along each chromosome, (2) defining a background noise model, (3) identifying candidate peak locations, and (4) assessing the significance of each candidate peak [36]. Peak-calling tools in earlier time quantify fold enrichment between samples of interest and expected background, and then apply the Poisson model to assess the significance of the enriched regions [36]. Recently developed peak-calling tools use the strand-dependent bimodality information and adopt a more realistic background model to capture local variations [167].

3.2.3 Gene Expression-based Predictive Modeling

Other than DEG detection, building prediction models using gene expression is also a popular RNA-seq application. Gene expression-based predictive modeling involves training classification models using gene expression tables with known clinical outcome as labels. Such the models can be used for future outcome prediction for any newly collected gene expression profiles. To assess prediction performance of certain clinical endpoints of interest without overfitting, the nested cross-validation technique is very popular that involves training and testing of an optimal prediction model. This is accomplished using the k-fold optimizing or inner cross-validation, applied to the training subset from the m-fold outer cross-validation. Once the final optimal prediction model parameters (i.e., the classifier hyperparameters and feature size) are identified, the final

model is trained using the entire training subset, and then tested using the remaining fold from the m-fold outer cross-validation. This process was repeated for several iterations to improve the robustness of prediction performance estimation. The nested cross-validation can be used with any kind of classifiers, such as adaptive boosting (AdaBoost), k-nearest neighbors (KNN), logistic regression (LR), random forests (RF), and support vector machines (SVM). In addition, since the number of genes (i.e., the number of features) is much larger than the number of samples, feature selection is necessary to reduce the dimensionality of the feature space. For gene expression data, the minimum redundancy, maximum relevance (mRMR) feature selection method is a popular method [168].

3.3 Case Study

To demonstrate my proposed approach for knowledge discovery using RNA-seq or ChIP-seq data for precision medicine, I detail two case studies in this section, including biomarker identification for cardiovascular diseases and predictive modeling for cancers. The background, experimental design, datasets, and results of the first case study (i.e., biomarker identification for cardiovascular diseases) are discussed based on its original publication [157], and those of the second case study (i.e., predictive modeling for cancers) are in preparation for submission to *Nature Methods*.

3.3.1 Biomarker Identification for Cardiovascular Diseases

3.3.1.1 Background

Cardiovascular disease (CVD) is the leading cause of death worldwide. Prediction and prevention of CVD, such as coronary artery disease and atherosclerosis, traditionally depend on identification of risk factors [169, 170]. These factors are effective in the

general assessment of CVD risk, but are not consistent indicators for all individuals [171]. Therefore, CVD research has recently been expanded to include the identification of -omic biomarkers (e.g., genomic, transcriptomic, and epigenomic) that may (1) improve the understanding of the molecular mechanisms of CVD, (2) facilitate the development of personalized CVD care, and (3) reduce CVD mortality rates by accurately identifying high-risk individuals [172]. NGS is a promising technology to identify -omic biomarkers. Because of its high-throughput capability in discovering novel genomic features with base-pair resolution, NGS is projected to play an increasingly important role in clinical diagnostics and personalized medicine for CVD [173, 174].

NGS and associated bioinformatics methods have been applied to cardiovascular genomics, transcriptomics, and epigenomics. **Figure 35** illustrates four NGS applications such as (A) identification of DEGs using RNA-seq, (B) identification of protein-binding regions in the genome using ChIP-seq, (C) identification of genetic variants in exon regions using exome sequencing, and (D) identification of genomic methylation patterns using methyl-CpG-binding domain sequencing (MBD-seq). These applications identify and quantify -omic biomarkers that may be clinically viable for early disease diagnosis and effective disease treatment and management. In this case study, I focus on two major applications of NGS technology: (1) RNA-seq, which has enabled researchers to characterize CVD by studying transcriptome-wide expression profiles [175], alternative splicing patterns [176], and miRNA regulatory networks [177]; and (2) ChIP-seq, which has enabled researchers to examine the epigenetic mechanisms of CVD by profiling the genome-wide pattern of protein-binding regions (e.g., transcription factors and enhancers) or histone modifications [178, 179].

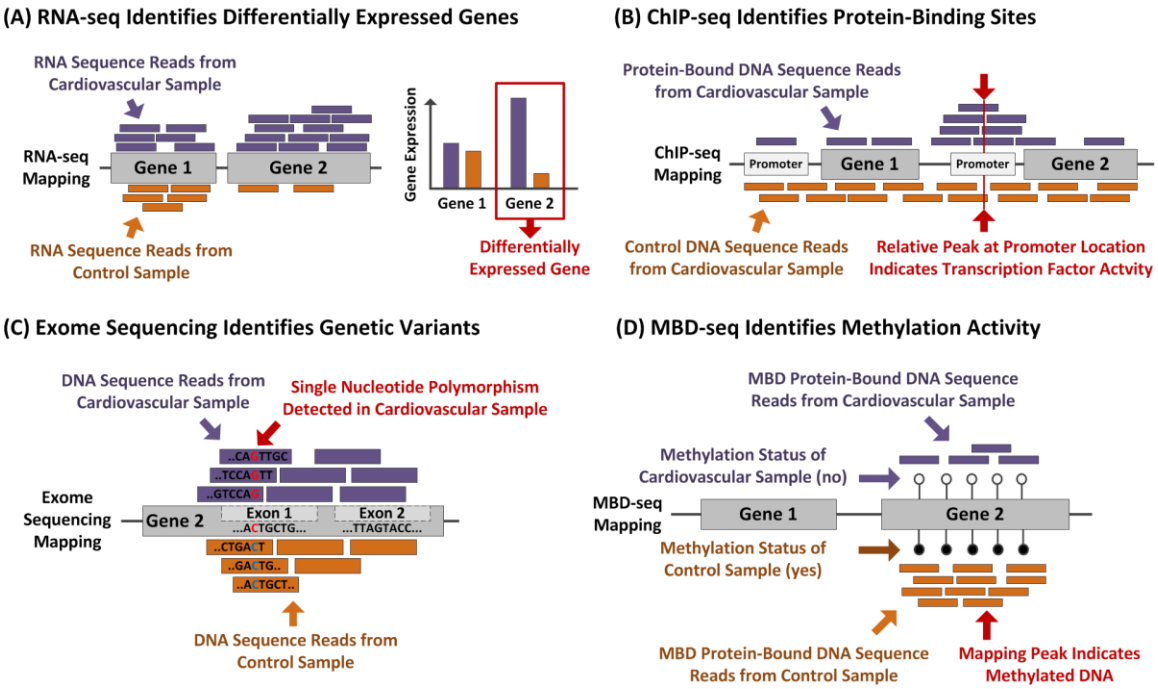


Figure 35: NGS Facilitates the Identification of -Omic Biomarkers for CVD. (A) RNA-seq detects differentially expressed genes by comparing gene expression profiles of CVD samples to those of control samples. (B) ChIP-seq identifies transcription factor activity by detecting peaks formed by mapping DNA sequence reads that bind to transcription factor proteins. Transcription factor activity correlates with gene expression. (C) Exome sequencing detects genetic variants such as SNPs that may correlate with CVD phenotypes. (D) MBD-seq is similar to ChIP-seq, but identifies regions of DNA methylation, which can affect gene expression.

I demonstrate the utility of NGS bioinformatics for cardiovascular research by applying RNA-seq and ChIP-seq pipelines to publicly available CVD RNA-seq and ChIP-seq datasets downloaded from the NCBI SRA repository. In each NGS application, I illustrate the NGS bioinformatics solution for CVD research; evaluate the performance of the critical step that identifies CVD biomarkers; and discuss the remaining bioinformatics challenges.

3.3.1.2 Experimental Design

DEG Detection Using RNA-seq

The bioinformatics pipeline for identifying DEGs using RNA-seq data includes sequence mapping, expression quantification, expression normalization, and DEG detection (**Figure 36**). I use the same sequence mapper TopHat [83] and expression quantifier Cufflinks [47] with eight different DEG detection tools to construct eight pipelines. Each DEG detection tool implements an expression normalization method that optimizes its DEG detection performance. TopHat maps four RNA-seq samples (detailed in the Section 3.3.1.3 "Datasets") to the GRCm38/mm10 mouse genome [180] with the guidance of the RefSeq genome annotation [125]. Cufflinks quantifies gene and transcript expression in terms of raw read counts. The eight DEG detection tools (**Figure 36**, left table) include both nonparametric (e.g., SAMseq [160] and NOISeq [161]) and parametric methods (e.g., baySeq [164], Cuffdiff2 [166], DESeq2 [98], DSS [181], edgeR [65], and Limma+Voom [182]) that represent a wide variety of prevalent and novel algorithms. The significant DEGs are identified with adjusted p-values less than 0.05. To evaluate the performance of the eight pipelines, I have designed five metrics:

(1) The authors of the original paper [183] performed qPCR validation for 16 cardiac hypertrophy- or fibrosis-related genes. Among 16 genes, only 12 were reported to be significantly differentially expressed. To assess the power of RNA-seq, I use the concordance of these 12 DEGs between qPCR and eight RNA-seq pipelines as the first metric. For genes detected by less than four out of the eight pipelines, I further investigated those expression patterns to understand the cause.

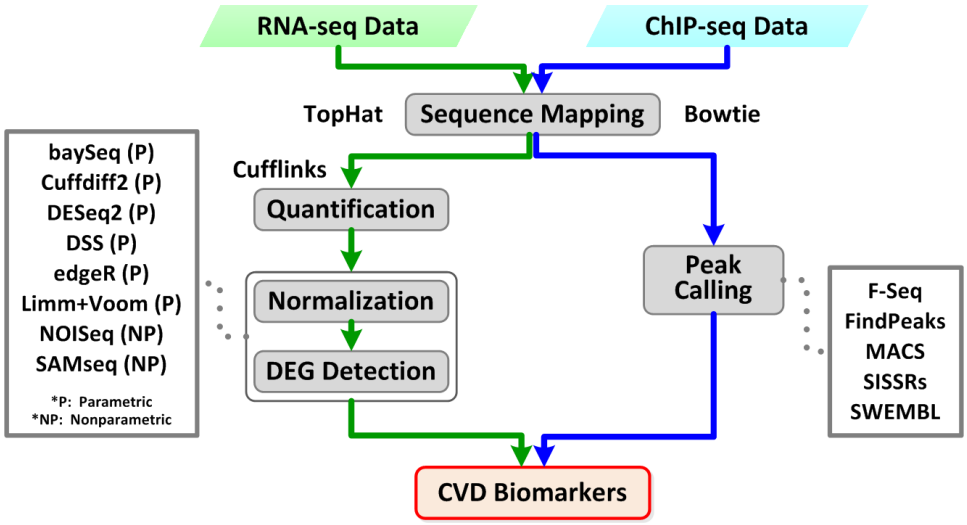


Figure 36: Bioinformatics Pipelines for RNA-seq and ChIP-seq Data. Green arrows indicate the pipeline for RNA-seq data and blue arrows represent the pipeline for ChIP-seq data. Selected DEG detection and peak-calling tools are listed following alphabetical order. For DEG detection tools, P stands for parametric and NP stands for nonparametric.

(2) To assess the biological relevance of DEGs, I use the ToppFun web-based tool in the ToppGene Suite [184] to annotate the functions of DEGs in terms of 114 significant Gene Ontology (GO) terms and four significant pathways with adjusted p-values less than 0.05. The GO terms and pathways associated with the 16 cardiac hypertrophy- or fibrosis-related genes are defined as the ground-truth functional annotation (**Figure 37**, Panel B). I use the concordance between the pipeline-specific annotations and the ground-truth as the second metric.

(3) To assess the reproducibility among various DEG detection tools, I compute the number of overlapping DEGs among the eight tools as the third metric.

(4) To assess the expression profiles of DEGs, I used the ratio of the dominant read count (i.e., the largest read count of a DEG across all samples) to the total read count for any DEG g, and its distribution as the fourth metric:

$$R_{dominance,g} = \frac{Max(A_{1,g}, A_{2,g}, B_{1,g}, B_{2,g})}{Sum(A_{1,g}, A_{2,g}, B_{1,g}, B_{2,g})}. \quad \textbf{(19)}$$

$A_{1,g}$, $A_{2,g}$, $B_{1,g}$, and $B_{2,g}$ are normalized read counts after adjusting the sequencing depth effect for samples A_1, A_2, B_1, and B_2 for any DEG g. [A_1, A_2] and [B_1, B_2] are biological replicates for the wild-type and Ezh2-deficient samples , respectively (detailed in the Section 3.3.1.3 "Datasets"). Given **Equation (19)**, the range of $R_{dominance,g}$ is from 25% (i.e., $A_{1,g}$=$A_{2,g}$=$B_{1,g}$=$B_{2,g}$) to 100% (i.e., *Max=Sum*) with a few possible scenarios: (a) if a gene is not significantly differentially expressed and the variability between replicates is small, the normalized read counts $A_{1,g}$, $A_{2,g}$, $B_{1,g}$, and $B_{2,g}$ will only differ slightly from one another, and the $R_{dominance,g}$ will be close to 25%; (b) if a gene is highly differentially expressed and the variability between replicates is small, $R_{dominance,g}$ will be around 50%; and (c) if the variability between replicates is large, $R_{dominance,g}$ can be significantly greater than 50% (e.g., $R_{dominance,g}$= 80% if [$A_{1,g}$, $A_{2,g}$, $B_{1,g}$, $B_{2,g}$] = [120, 30, 0, 0]).

(5) To assess the capability of each tool for detecting highly-expressed or low-expressed DEGs, I calculate the mean read count of each DEG g from the normalized read counts (i.e., $A_{1,g}$, $A_{2,g}$, $B_{1,g}$, and $B_{2,g}$) as the fifth metric.

Peak Calling Using ChIP-seq

The bioinformatics pipeline for identifying genome-wide protein-binding regions using ChIP-seq includes sequence mapping and peak calling (**Figure 36**). I use the same sequence mapper, Bowtie, and six different peak-calling tools to construct totally six pipelines. Bowtie [73] maps sequence reads (or sequence "tags" in ChIP-seq) to the GRCh37 / hg19 human genome [185] and reports only uniquely mapped tags. The six

peak-calling tools (**Figure 36**, right table), including SISSRs [57], MACS [56], FindPeaks [186], SWEMBL [187], SICER [188], and F-Seq [189] represent a wide variety of algorithms for determining statistically significant peaks. I run these tools using their default or recommended parameters with a p-value threshold of 10^{-3}. The identified peaks are putative protein-binding regions for p300 and CBP proteins (detailed in the Section 3.3.1.3 "Datasets").

To assess the performance of the six peak-calling tools, I have designed five metrics: (1) to visualize sequence-mapping and peak-calling information using the Integrative Genomics Viewer [190]; (2) to count the total number of peaks called by each tool; (3) to investigate the distribution of N_i, the normalized tags per peak, as defined in the **Equation (20)**:

$$N_i = \frac{(Number\ of\ Tags)_i}{(Peak\ Length/100)_i}; \quad \textbf{(20)}$$

(4) to compute the average length of peaks called by each tool, as defined in the **Equation (21)**:

$$L_{avg} = \frac{\sum_{i=1}^{N} Length(Peak_i)}{N}, \quad \textbf{(21)}$$

where N is the total number of peaks; and (5) to biologically validate the peaks by investigating the percentage of peaks that contain at least one p300 motif using the FIMO (find individual motif occurrences) program [191] with a p-value threshold of 10^{-4}. The input information for the FIMO program includes DNA sequences corresponding to these peaks and the position-specific scoring matrix for the p300 motif retrieved from the SwissRegulon Portal [192].

3.3.1.3 Datasets

The RNA-seq dataset (SRA accession: SRP009662) was acquired to investigate the effects of Ezh2 deletion on postnatal cardiac development, homeostasis, and gene expression [183]. The authors reported that the loss of Ezh2 gene in cardiac precursors would lead to cardiac hypertrophy and fibrosis. This dataset contains wild-type and Ezh2-deficient adult mouse right ventricle samples, each with two biological replicates. Each sample was sequenced with the Illumina HiSeq 2000 platform and contains around 30 million 2×50 bp read pairs.

The ChIP-seq dataset (SRA accession: SRP008658) investigated the genome-wide map of human heart enhancers with a pan-specific antibody that targets two closely-related transcriptional coactivator proteins, p300 and CBP (CREB-binding protein) [193]. This dataset contains tissue samples from one fetal and one adult human heart. Each sample was sequenced with Illumina Genome Analyzer and contains around 27 million 36 bp single-ended reads.

3.3.1.4 Results and Discussion

DEG Detection Using RNA-seq

I have evaluated the performance of the eight DEG detection tools by five metrics (**Figure 37**, Panels A-F). Using the 12 qPCR-validated DEGs, Panel A shows that nonparametric methods such as NOISeq and SAMseq, and parametric methods such as Cuffdiff2 and edgeR, were able to identify at least half of these 12 DEGs. In contrast, parametric methods, such as baySeq, DESeq2, DSS, and Limma+Voom, were able to identify only one or two of these 12 DEGs. Six genes were difficult to detect by RNA-seq-based methods (marked in red). The expression pattern of these difficult genes shows

that they have either smaller fold changes (e.g., Tgfb3) or higher between-replicate variability (e.g., Actn3).

Panel B listed the top 20 GO terms and all four pathways from the ground-truth functional annotations, most of which were linked to the mechanisms of muscle contraction and heart development. Panel C summarized the concordance between the pipeline-specific annotations and the ground-truth in terms of the top 20 GO terms (ranked by p-values), all GO terms, and all pathways. DEGs detected by baySeq, DSS, and Limma+Voom were associated with zero GO terms and only a few pathways, which suggested that these tools detected DEGs with very diverse functions. DEGs detected by edgeR and Cuffdiff2 had more high-rank functional annotations concordant with the ground-truth annotation. In contrast, DEGs detected by DESeq2, SAMseq, and NOISeq were linked to many GO terms and pathways that were biologically irrelevant to the original study, with no concordance appeared in the top 20 GO terms.

Panel D showed the number of DEGs supported by one, two, or all eight tools for each DEG detection method. baySeq, DSS, and Limma+Voom identified a fewer number of DEGs (i.e., 32, 23, and 12, respectively) that were highly reproducible among various tools (i.e., each DEG were supported by at least two other tools). Tools with more detected DEGs, such as SAMseq, NOISeq, and DESeq2, tended to have more pipeline-specific or unique DEGs. However, as discussed earlier, a higher number of DEGs did not necessarily lead to more biologically relevant results. Panel E demonstrated the distribution of $R_{dominance,g}$ for DEGs. Most DEGs had $R_{dominance,g}$ in the range of 25% to 60% following the scenarios (4a) and (4b) I have discussed in Section 3.3.1.2 "Experimental Design" for RNA-seq. Such observation indicated that most DEGs

detected by RNA-seq pipelines did not have huge variability between biological replicates. DEGs with larger between-replicate variability resulted in $R_{dominance,g}$ greater than 60%. Such high variability can be the nature of biological replicates or biases introduced in the sequencing or analytical processes. The nonparametric NOISeq method had the highest percentage of DEGs with $R_{dominance,g}$ greater than 60% since it identified many genes with very low read counts (e.g., $[A_{1,g}, A_{2,g}, B_{1,g}, B_{2,g}]=[1, 1, 0, 0]$). Thus, a small deviation in the read counts may have caused a huge variation in $R_{dominance,g}$. For parametric methods, higher $R_{dominance,g}$ indicated that the read counts may not follow a negative binomial distribution [160]. edgeR and Cuffdiff2 had a higher chance of detecting this type of genes as DEGs. Panel F showed the distribution of the mean read counts of DEGs. NOISeq had a bimodal distribution because of its tendency to identify some DEGs with very low read counts. The other seven tools shared a similar range of the mean read counts of DEGs, with baySeq slightly skewed to the left (i.e., lower mean read counts).

(A)

qRT-PCR Validated DEGs

Tool (#DEG)	baySeq (32)	Cuffdiff2 (155)	DESeq2 (514)	DSS (23)	edgeR (243)	Limma+ Voom (12)	NOISeq (792)	SAMseq (2401)	Expression Pattern			
									A1	A2	B1	B2
Nppa		V			V	V	V	V	81596	57346	9593	3985
Nppb		V			V		V	V	9201	6232	2815	1161
Myh7		V			V		V	V	30488	17510	1822	7235
Tgfb3									675	393	523	267
Postn		V	V		V		V	V	1717	1252	320	539
Spp1								V	33	40	4	21
Six1	V		V	V	V		V	V	116	108	14	3
Eya1								V	46	33	11	19
Myh8							V	V	4	3	0	0
Myl1		V			V		V	V	606	1536	199	128
Myl4		V						V	631	500	332	35
Actn3					V		V	V	94	4	2	0
Total	1/12	6/12	2/12	1/12	7/12	1/12	8/12	11/12				

(B)

Gene Ontology Term	Adjusted P-Value	Gene Ontology Term	Adjusted P-Value
GO:0030049 Muscle filament sliding	1.33E-14	GO:0015629 Actin cytoskeleton	1.12E-09
GO:0033275 Actin-myosin filament sliding	1.33E-14	GO:0016459 Myosin complex	1.12E-09
GO:0070252 Actin-mediated cell contraction	9.77E-13	GO:0006936 Muscle contraction	1.20E-09
GO:0008307 Structural constituent of muscle	2.90E-12	GO:0030016 Myofibril	1.08E-08
GO:0030048 Actin filament-based movement	5.89E-11	GO:0007507 Heart development	2.54E-07
GO:0003012 Muscle system process	5.89E-11	GO:0030017 Sarcomere	2.65E-07
GO:0044449 Contractile fiber part	5.37E-10	GO:0006941 Striated muscle contraction	2.81E-07
GO:0005859 Muscle myosin complex	5.37E-10	GO:0003007 Heart morphogenesis	9.56E-07
GO:0043292 Contractile fiber	7.10E-10	GO:0003779 Actin binding	4.27E-06
GO:0016460 Myosin II complex	1.12E-09	GO:0061061 Muscle structure development	4.97E-06

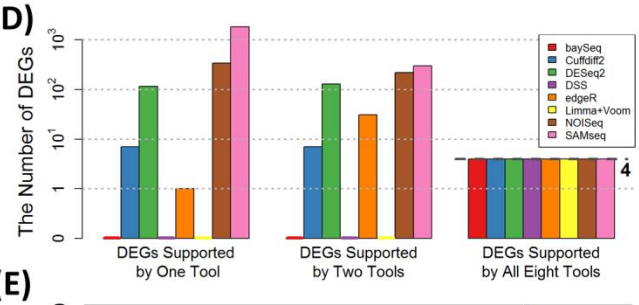
Pathway	Adjusted P-Value
Reactome: Genes involved in Striated Muscle Contraction	5.33E-12
WikiPathways: Striated Muscle Contraction	1.27E-11
Reactome: Genes involved in Muscle contraction	7.72E-11
BioCarta: ALK in cardiac myocytes	2.85E-05

(C)

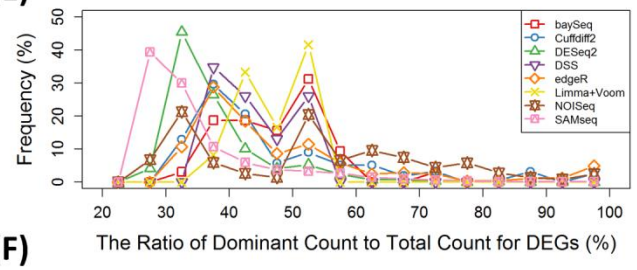
Tools (#DEGs)	Number of Significant GO Terms*	Number of Concordant GO Terms between Ground Truth and Each Tool (Top 20)	Number of Concordant GO Terms between Ground Truth and Each Tool (All)**	Number of Significant Pathways*	Number of Concordant Pathways between Ground Truth and Each Tool (All)***
baySeq (32)	0	0	0	0	0
Cuffdiff2 (155)	107	8	28	18	2
DESeq2 (514)	93	0	7	20	0
DSS (23)	0	0	0	2	0
edgeR (243)	75	12	32	17	3
Limma+ Voom (12)	0	0	0	0	0
NOISeq (792)	171	0	33	28	3
SAMseq (2401)	406	0	25	74	3

* False Discovery Rate = 0.05
** Ground truth contains 114 significant GO terms
*** Ground truth contains 4 significant pathways

(D)



(E)



(F)

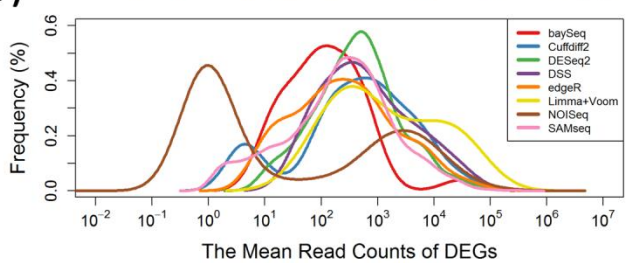


Figure 37: Functional and Quantitative Assessment of DEG Detection Tools. (A) Checkmarks indicated concordant DEGs between qPCR validation conducted by the original study and the results from the eight DEG detection tools. Genes marked in red were identified by less than four out of the eight tools. (B) The top 20 significant GO terms and all four significant pathways of the ground truth functional annotation, which was established by annotating the 16 qPCR validated genes. (C) The number of concordant GO terms and pathways between ground-truth and pipeline-specific annotations. (D) The number of DEGs supported by one, two, or all eight tools for each DEG detection pipeline. (E) The distribution of the ratio of dominant read counts to total read counts of all DEGs for each DEG detection pipeline. (F) The distribution of the mean read counts of all DEGs for each DEG detection pipeline.

Peak Calling Using ChIP-seq

I have investigated the performance of the six peak-calling tools by five metrics (**Figure 38**, Panels A-E). Visualizing by the Integrative Genomics Viewer, Panel A showed the peak regions called by the six tools with corresponding coverage information from the Bowtie alignment in the upstream region of the INPP5A gene (inositol polyphosphate-5-phsphatase, 40kDa). The selective nature of SISSRs and MACS resulted in sparse and short peaks. In contrast, FindPeaks and SWEMBL tended to call very long peaks with lengths over 10 kbp. Panel B demonstrated the number of peaks called by the six tools. SICER called the largest number of peaks, followed by FindPeaks and F-Seq. SICER failed to form longer peaks by merging nearby peaks, resulting in a relatively higher number of peaks. FindPeaks called two separate peaks even though two protein-binding regions were in close proximity; thus, FindPeaks also tended to call more peaks than the other tools. Panel C depicted the distribution of the number of tags per peak normalized by the peak length. Larger numbers indicated that the detected peaks were supported by more evidence. SISSRs, MACS, SWEMBL, and F-Seq exhibited a moderate to high number of tags per peak. In contrast, FindPeaks and SICER detected

some peaks with a very low number of tags per peak. These peaks may not have been reliable because of limited evidence. Panel D showed the average length of peaks called by the six tools. Among them, SWEMBL had the longest average length, which may not have been a reasonable length for protein DNA-binding sites. SISSRs, MACS, and F-Seq exhibited the average peak length of less than 400 bp, which was close to the designed fragment length from the Illumina sequencing protocol.

Using the FIMO program, Panel E demonstrated the percentage of peaks that contained the p300 motif. MACS performed the best with 15% to 23% of the peaks containing the motif. SISSRs and F-Seq performed moderately well with their motif discovery rate ranging from 6% to 8%. Even though FindPeaks and SICER detected a significantly larger number of peaks than the others, only 2% to 3% of these peaks contained the p300 motif, exposing their relatively high false positive rates. Around 11% to 22% of peaks called by SWEMBL contained the p300 motif. However, despite such high performance, the peaks were not reliable since SWEMBL had extremely long peaks on average, which increased the probability of identifying the motif by chance alone.

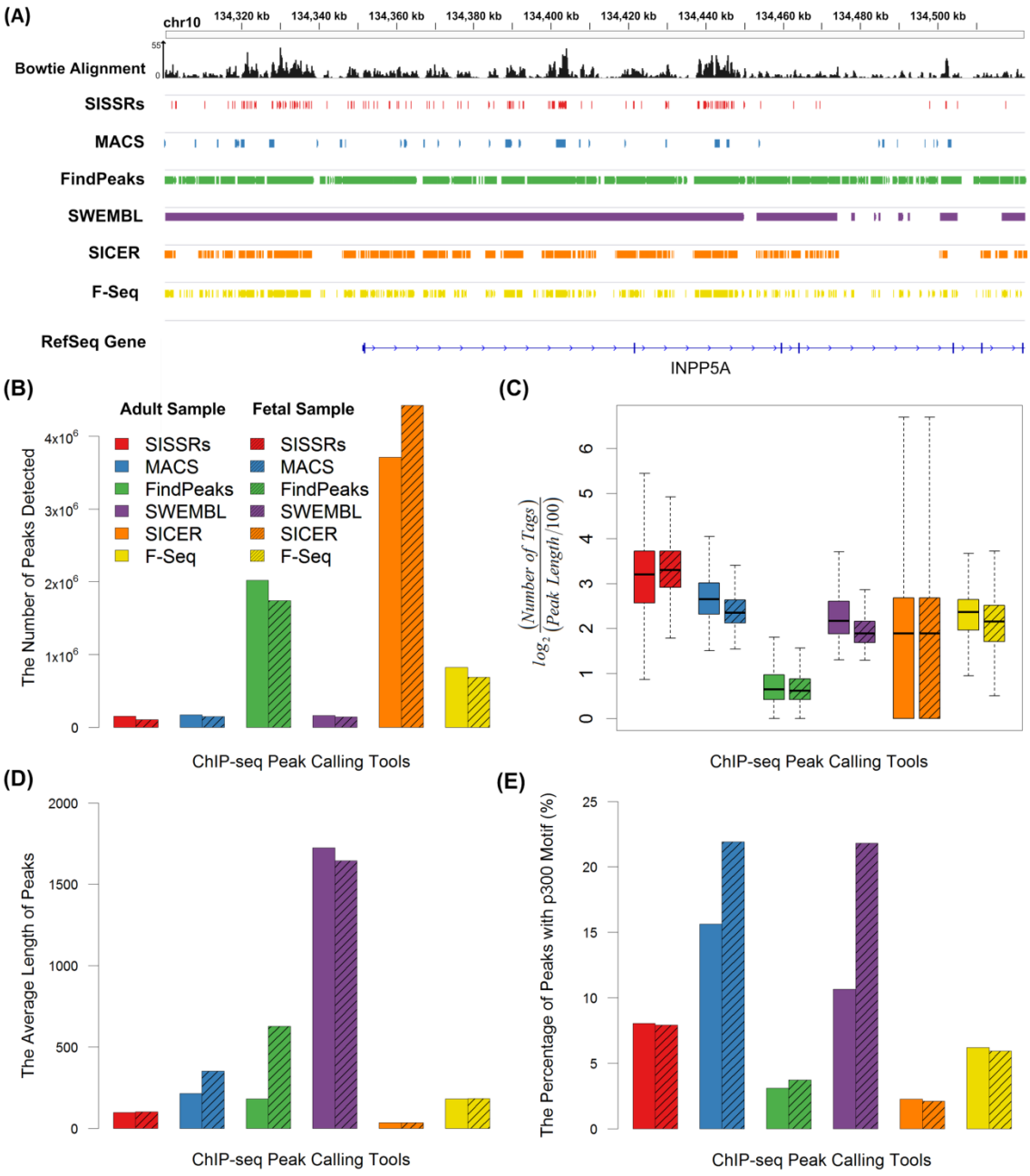


Figure 38: Qualitative and Quantitative Assessment of Various Peak-Calling Tools. (A) IGV visualized peaks called by the six tools in the upstream region of the INPP5A gene using the adult heart sample. The black histogram on the top represented the coverage of Bowtie alignment in the same region. (B) The number of peaks detected by each peak-calling pipeline. (C) The distribution of the number of tags per peak normalized by the peak length for each peak-calling pipeline. (D) The average length of peaks for each peak-calling pipeline. (E) The percentage of peaks that contains at least one p300 motif identified by the FIMO program with a p-value threshold of 10^{-4} for each peak-calling pipeline.

3.3.1.5 Summary of Case Study

In summary, the original paper of the RNA-seq dataset studied the effect of Ezh2 deletion on gene expression profiles. It identified a set of DEGs relevant to cardiac tissue development and remodeling [183]. My study examined the functions of DEGs detected by the eight RNA-seq pipelines, and edgeR and Cufflinks yielded the most functionally relevant DEGs. The nonparametric methods such as NOISeq and SAMseq identified many more DEGs than other tools, yet a large proportion of these DEGs may have been less reliable (e.g., DEGs with very low read counts) and irrelevant to the biology of the original study. RNA-seq technology provides an opportunity to comprehensively study the transcriptome. While fixing sequence mapping and expression quantification steps and focusing on evaluating only DEG detection methods, I found that different tools generated very different DEG sets. Therefore, translating the computational findings into real clinical applications requires integrative biological interpretation and large-scale experimental validation. In addition, RNA-seq technology can be unreliable for estimating expression of low-expressing genes, but currently no standardized methods are capable of handling them properly. Thus, distinguishing true signals from noise for low-expressing genes remains a challenge.

For the ChIP-seq dataset, the original study used ChIP-seq with the antibody that recognizes the enhancer-associated coactivator proteins p300 and CBP to annotate candidate heart enhancers that may regulate the expression of heart development-related genes in the human genome. By examining the percentage of peak regions (i.e., candidate heart enhancer regions) that contained the p300 motif, MACS achieved the highest motif discovery rate among the six tools, which suggested that MACS identified more

biologically relevant peaks than the others. In addition, MACS's peaks had the second highest tag coverage and the reasonable average peak length. In contrast, SICER identified peaks with the lowest motif discovery rate and very low tag coverage. Most peak-calling tools need control samples for building background models essential for conducting statistical tests. These background models can be local or global. The global model is easier to build but lacks the consideration of local biases. Accurately identifying peaks requires an adaptive background signal model that can dynamically change parameters to accommodate local variations and different ChIP-seq experiments.

3.3.2 Prediction Models for Cancers

3.3.2.1 Background

This case study is part of the SEQC project, and it is also the continuation of Case Study 4 in Chapter 2 (i.e., Section 2.4.4). Please refer to Section 2.4.4 for detailed information about the RNA-seq pipeline study focusing on benchmark datasets and associated benchmark metric performance. In contrast to Chapter 2, Case Study 4, this case study examines the effect of RNA-seq pipelines on downstream gene expression-based prediction of disease outcome. Similar to Chapter 2, Case Study 4, we did a comprehensively literature survey about the effect of RNA-seq pipelines on downstream prediction performance. To the best of our knowledge, *no studies have reported the effect of RNA-seq pipeline components on gene-expression-based prediction performance and no guidelines exist for selecting RNA-seq pipelines for prediction of disease outcome.*

The FDA coordinated with the BGI to generate a clinical dataset consisting of neuroblastoma patient samples (referred to as SEQC-neuroblastoma) [194], and then provided this datasets to SEQC teams to investigate the joint impact of pipeline

components on downstream gene expression-based prediction. We quantified gene expression in the SEQC-neuroblastoma dataset and the lung adenocarcinoma dataset (referred to as TCGA-lung-adenocarcinoma) downloaded from The Cancer Genome Atlas (TCGA), and then determining if RNA-seq pipeline components contributed to variations in prediction performance of disease outcome.

3.3.2.2 Experimental Design

We used the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets to assess the effect of upstream RNA-seq pipeline components on downstream prediction of disease outcome. The total number of RNA-seq pipelines considered was 278, as summarized in **Table 12**. The SEQC-neuroblastoma dataset, provided by the SEQC consortium, contains RNA-seq data of 176 primary neuroblastomas obtained from high-risk patients with well-annotated clinical data [194], in which survival information, including event-free survival (EFS) and overall survival (OS), was used for defining group labels for predictive modeling. The TCGA-lung-adenocarcinoma dataset contains RNA-seq data of patients with known survival time used for defining group labels.

3.3.2.3 Datasets

We used a 176-sample neuroblastoma dataset (a subset of a larger 498-sample dataset; accession GSE47792) to assess the performance of RNA-seq pipelines in terms of gene expression-based prediction of disease outcome. These samples were provided by the University Children's Hospital of Cologne and sequenced at BGI using the Illumina platform [194]. All 176 samples were taken from high-risk patients that were defined as those either with stage 4 neuroblastoma and age >18 months or with MYCN-amplified tumors of any stage or age.

We predicted two clinical endpoints—event-free survival (EFS), that is, the occurrence of events such as progress, relapse, or death, and overall survival (OS), that is, death. For both endpoints, patients were partitioned into two groups (i.e., high risks versus long risks). High-risk patients experienced an event or died before a predefined survival-time threshold, while low-risk patients experienced an event or died after the threshold, or their last follow-up exceeded the threshold. Survival-time thresholds for EFS and OS were two and three years, respectively. The thresholds were chosen to balance the number of high-risk and low-risk patients. Details of the SEQC-neuroblastoma dataset are provided in **Table 16**.

Table 16: Prediction Endpoints for the SEQC Neuroblastoma Dataset.

Endpoint	Grouping Criteria	Number of Samples
Event-Free Survival (EFS) [Threshold = 2 years]	Event occurred after the threshold OR Patient's last follow-up exceeded the threshold (no information about patient's event occurrence after the last follow-up)	67
	Event occurred before the threshold	97
Overall Survival (OS) [Threshold = 3 years]	Patient died after the threshold OR Patient's last follow-up exceeded the threshold (no information about patient's survival after the last follow-up)	83
	Patient died before the threshold	70

We also used an 87-sample lung adenocarcinoma RNA-seq dataset from The Cancer Genome Atlas (TCGA) repository. The prediction endpoint was also survival, and we used the same criteria to define high-risk and low-risk groups with the survival-time threshold of two years. The two-year threshold was chosen to balance the number of

high-risk and low-risk patients. Details of the TCGA-lung-adenocarcinoma dataset are provided in **Table 17**.

Table 17: Prediction Endpoint for the TCGA Lung Adenocarcinoma Dataset.

Endpoint	Grouping Criteria	Number of Samples
Survival [Threshold = 2 years]	Patient died after the threshold OR Patient's last follow-up exceeded the threshold (no information about patient's survival after the last follow-up)	47
	Patient died before the threshold	40

3.3.2.4 Results and Discussion

We used the same set of 278 RNA-seq pipelines to process the SEQC-neuroblastoma dataset (we used only 156 out of the 278 pipelines for the TCGA-lung-adenocarcinoma dataset). For each set of estimated gene expression (278 sets for neuroblastoma and 156 sets for lung adenocarcinoma), we performed nested cross-validation (**Figure 39**) using three classifiers—AdaBoost, LR, and SVM. For each clinical endpoint—neuroblastoma EFS, neuroblastoma OS, and lung adenocarcinoma survival—we calculated the AUC (area under the receiver operating characteristics curve) and MCC (Matthews correlation coefficient), and visualized these using heatmaps (**Figure 40**, **Figure 41**, and **Figure 42**).

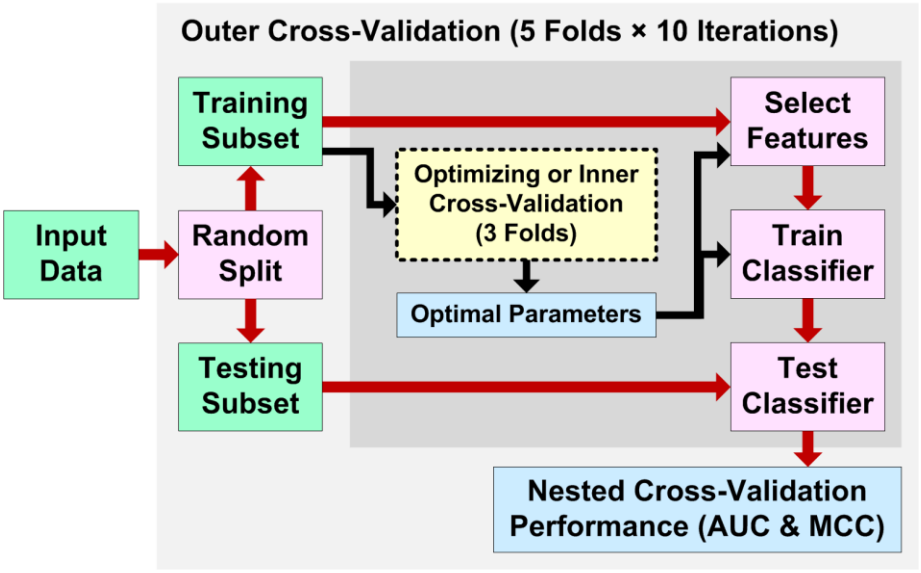


Figure 39: Predictive Modeling Using the Nested Cross-Validation Technique. For the outer cross-validation, input data are randomly split into training and testing subsets (green boxes) following the standard 5-fold cross-validation protocol. For each of the five training subsets, the 3-fold optimizing or inner cross-validation (yellow boxes) is applied to optimize the feature size and hyperparameters for classifiers. The optimal feature size and hyperparameters (blue boxes) are used to train a final classifier (pink boxes) that will be directly applied to the testing subset. The final predictive modeling performance is measured by both the area under the receiver operating characteristic curve (AUROC, or simply AUC) and the Matthews correlation coefficient (MCC).

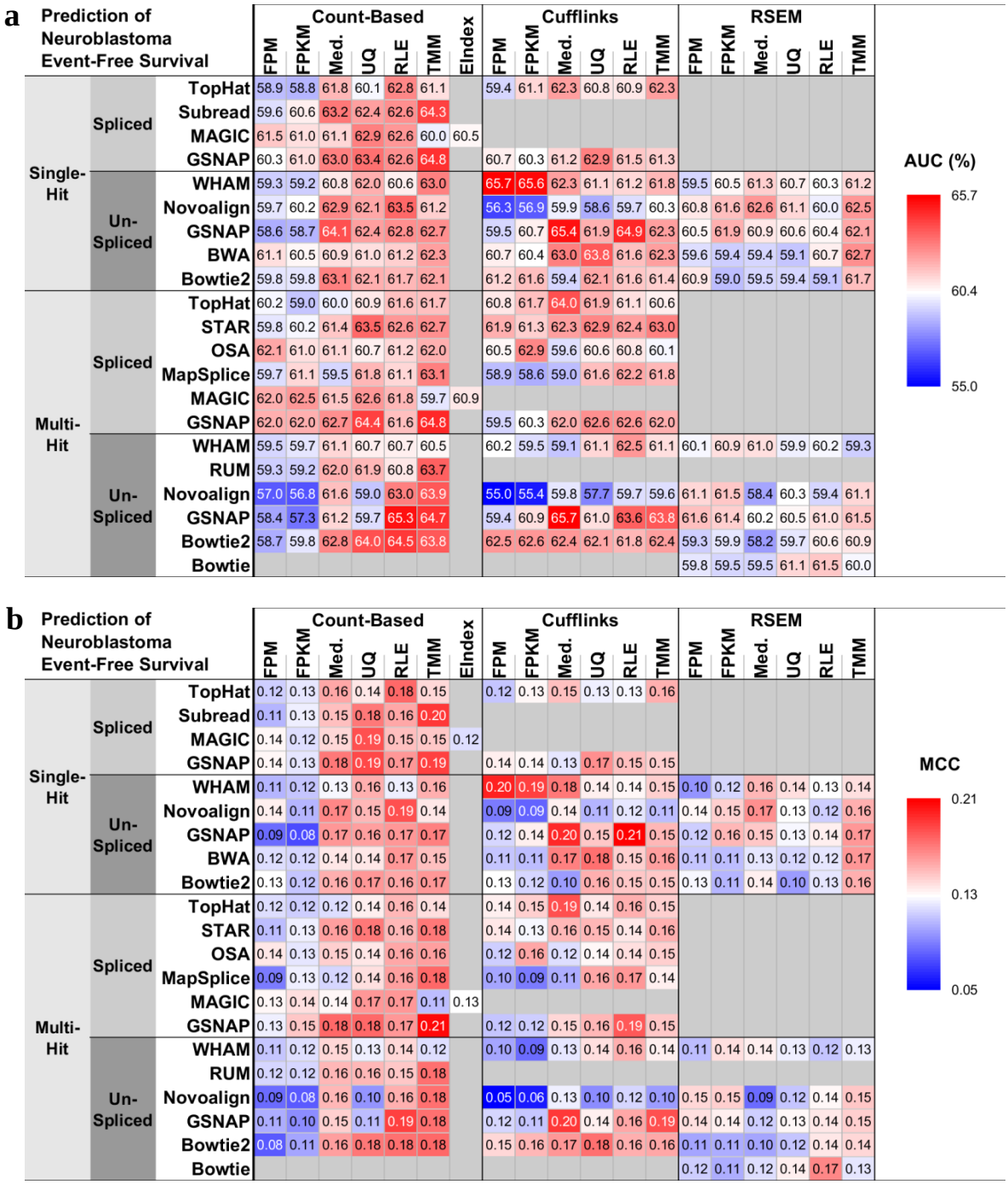


Figure 40: Prediction Performance of NB EFS Measured by AUC and MCC. The 278 RNA-seq pipelines applied to the SEQC-neuroblastoma (NB) dataset differ in terms of prediction performance measured by (a) AUC and (b) MCC. The predictive modeling procedure is detailed in **Figure 39**, and the prediction endpoint is dichotomized event-free survival (EFS) with the survival-time threshold of two years. Prediction performance is encoded as color, with red representing the highest AUC or MCC.

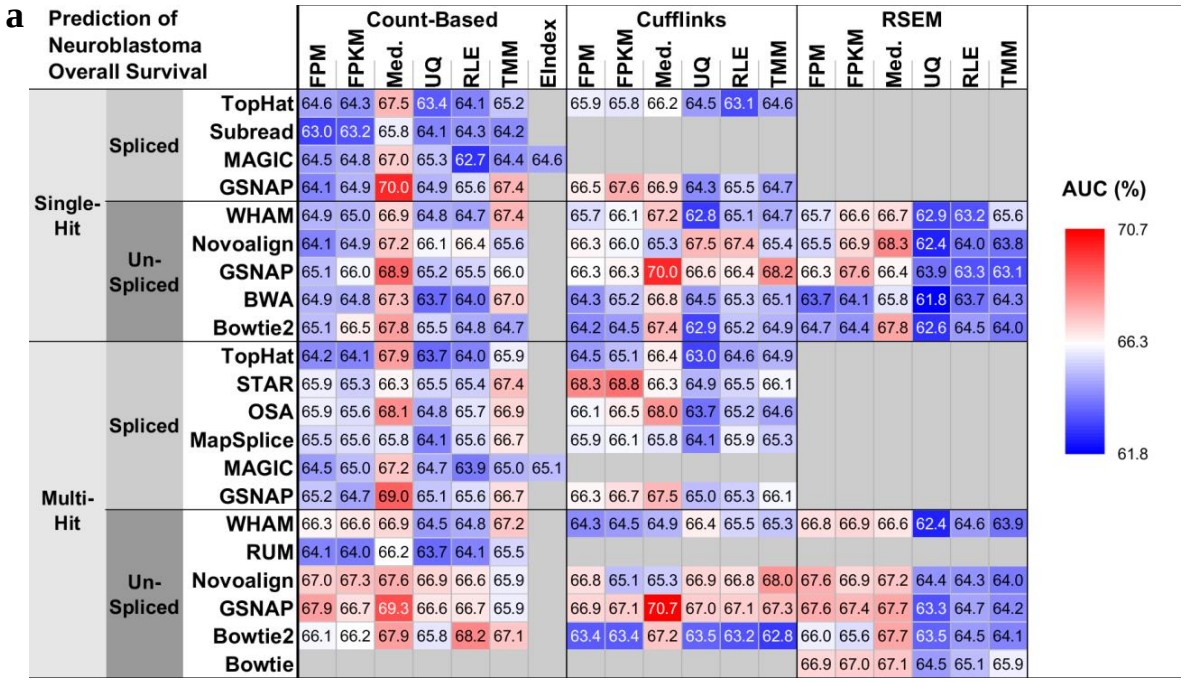


Figure 41: Prediction Performance of NB OS Measured by AUC and MCC. The 278 RNA-seq pipelines applied to the SEQC-neuroblastoma (NB) dataset differ in terms of prediction performance measured by (a) AUC and (b) MCC. The predictive modeling procedure is detailed in **Figure 39**, and the prediction endpoint is dichotomized overall survival (OS) with the survival-time threshold of three years. Prediction performance is encoded as color, with red representing the highest AUC or MCC.

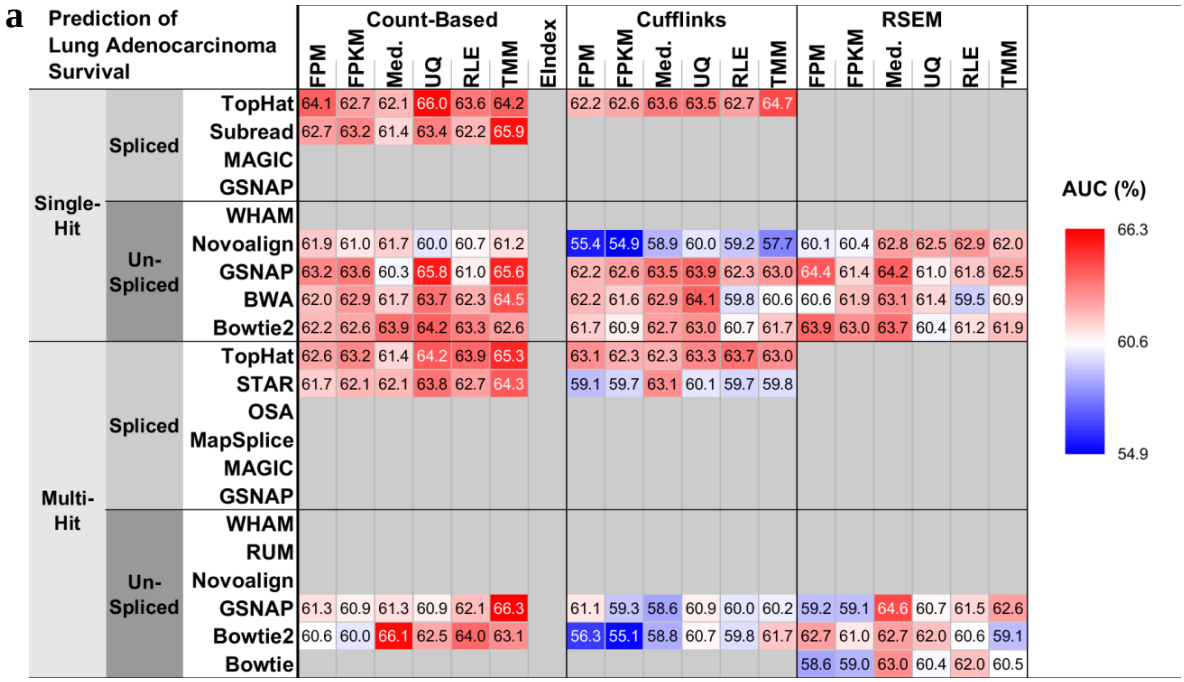


Figure 42: Prediction Performance of LUAD Survival Measured by AUC and MCC. The 156 RNA-seq pipelines applied to the TCGA-lung-adenocarcinoma (LUAD) dataset differ in terms of prediction performance measured by (a) AUC and (b) MCC. The predictive modeling procedure is detailed in **Figure 39**, and the prediction endpoint is dichotomized survival with the survival threshold of two years. Prediction performance is encoded as color, with red representing the highest AUC or MCC.

We observed the following results:

(1) For the neuroblastoma EFS endpoint, pipelines using count-based quantification with TMM, RLE, upper quartile, or median normalization tended to achieve high AUC and MCC; while those with FPM or FPKM normalization tended to perform poorly. In addition, Novoalign with Cufflinks and Bowtie2 or BWA with RSEM led to poor AUC and MCC, especially when combining with FPM or FPKM normalization (**Figure 40**).

(2) For the neuroblastoma OS endpoint, median normalization led to higher AUC and MCC than other normalization methods for most mapping-quantification combinations. GSNAP un-spliced mapping performed well with count-based or Cufflinks quantification but not RSEM quantification. In addition, pipelines with RSEM quantification and any of upper quartile, RLE, or TMM normalization tended to result in poor AUC and MCC (**Figure 41**).

(3) For the lung adenocarcinoma survival endpoint, pipelines with count-based quantification and TMM normalization tended to achieve high AUC and MCC. TopHat alignment with either count-based or Cufflinks quantification also performed well. In contrast, pipelines with any of Novoalign single-hit, STAR, GSNAP un-spliced multi-hit, or Bowtie2 multi-hit and Cufflinks resulted in lower AUC and MCC (**Figure 42**).

(4) ANOVA for each neuroblastoma endpoint showed that normalization was the largest statistically significant ($p<0.05$) source of variation, followed by mapping algorithm, two-way [mapping algorithm*quantification] interaction, and two-way [quantification*normalization] interaction (**Figure 43** and

Figure 44). For the lung adenocarcinoma endpoint, several pipeline components and their interactions contributed more evenly to the overall variance that may be due to only 156 pipelines were conducted (**Figure 45**). All ANOVA reported large residual variance that should be explained by higher order interactions.

Results suggested that the choice of upstream RNA-seq pipeline components significantly impacted the performance of downstream prediction of disease outcome.

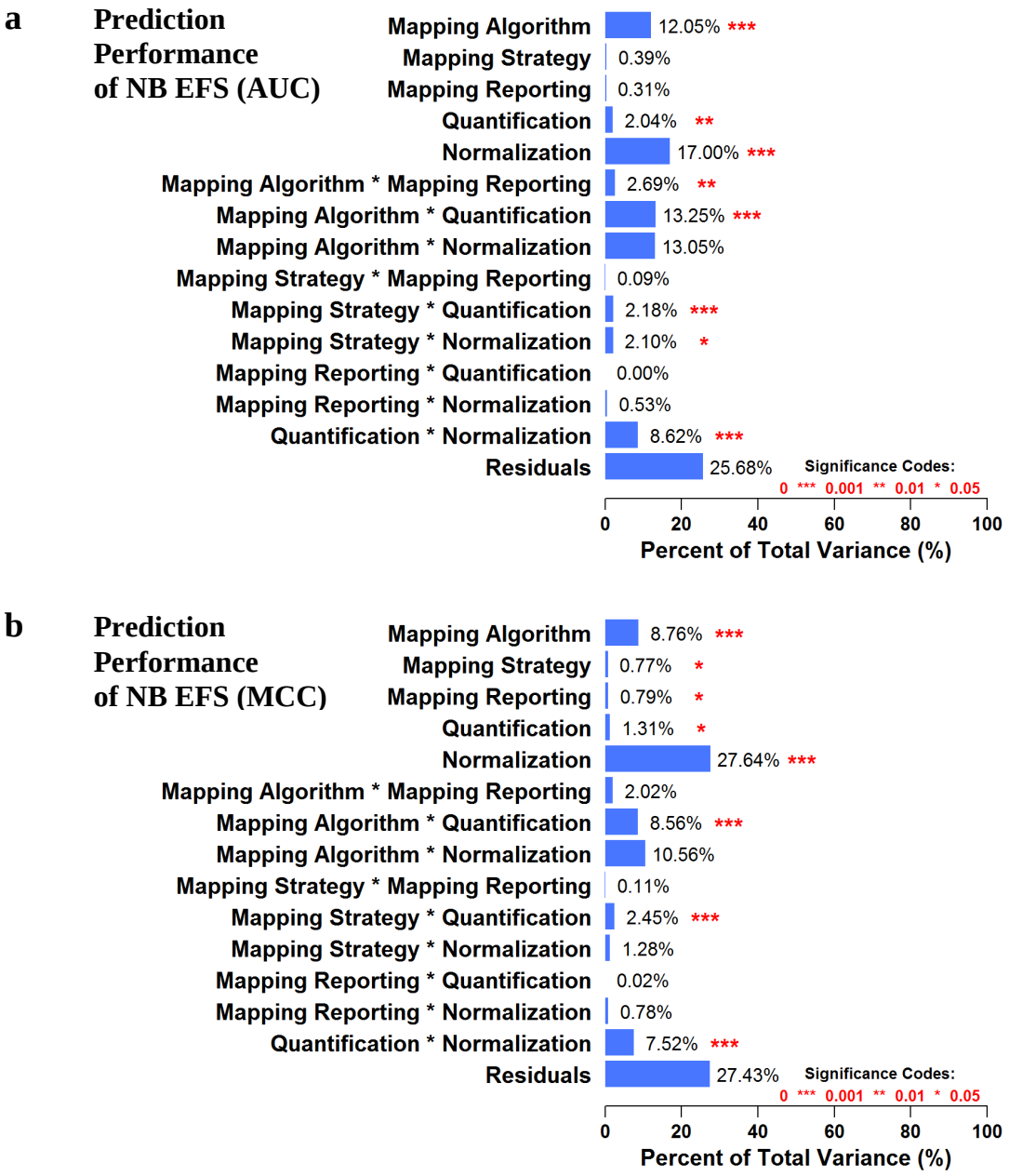


Figure 43: ANOVA for Prediction Performance of NB EFS. Analysis of variance (ANOVA) decomposes the overall variance in prediction performance of NB EFS into various factors considered, including RNA-seq pipeline components and associated two-way interactions. Panels (a) and (b) show the ANOVA for prediction AUC and MCC, respectively. Prediction performance of various classifiers has been averaged before applying the ANOVA. The statistical significance of the contribution of each component and interaction is denoted by red asterisks, with '***' indicates p-values are smaller than 0.001, '**' indicates p-values are smaller than 0.01, and '*' indicates p-values are smaller than 0.05. Among all components and interactions, the normalization contributes the most to the overall variance. Around a quarter of the overall variance belongs to residuals that cannot be explained by the factors considered.

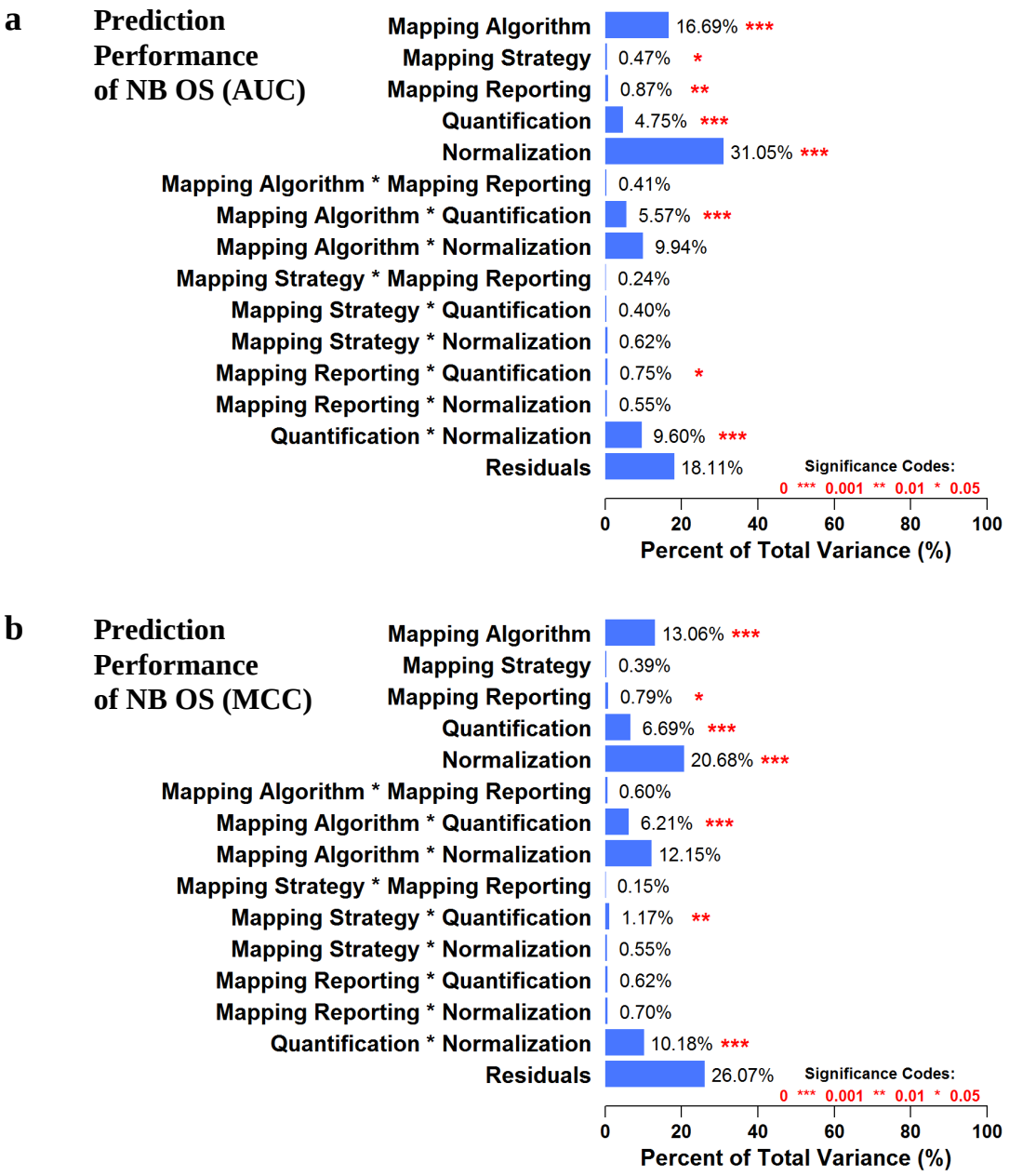


Figure 44: ANOVA for Prediction Performance of NB OS. Analysis of variance (ANOVA) decomposes the overall variance in prediction performance of NB OS into various factors considered, including RNA-seq pipeline components and associated two-way interactions. Panels (a) and (b) show the ANOVA for prediction AUC and MCC, respectively. Prediction performance of various classifiers has been averaged before applying the ANOVA. The statistical significance of the contribution of each component and interaction is denoted by red asterisks, with '***' indicates p-values are smaller than 0.001, '**' indicates p-values are smaller than 0.01, and '*' indicates p-values are smaller than 0.05. Among all components and interactions, the normalization contributes the most to the overall variance. Around a quarter of the overall variance belongs to residuals that cannot be explained by the factors considered.

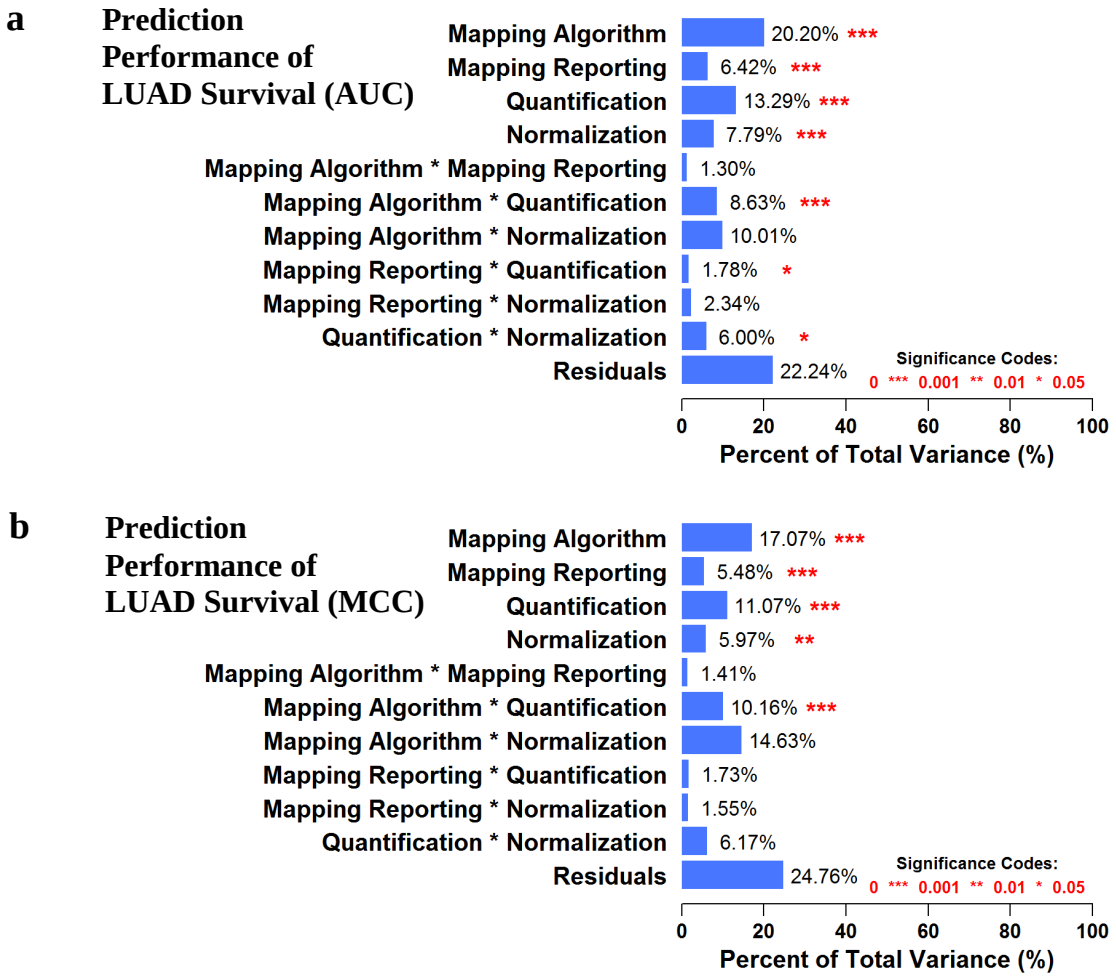


Figure 45: ANOVA for Prediction Performance of LUAD Survival. Analysis of variance (ANOVA) decomposes the overall variance in prediction performance of LUAD survival into various factors considered, including RNA-seq pipeline components and associated two-way interactions. Panels (a) and (b) show the ANOVA for prediction AUC and MCC, respectively. Prediction performance of various classifiers has been averaged before applying the ANOVA. The statistical significance of the contribution of each component and interaction is denoted by red asterisks, with '***' indicates p-values are smaller than 0.001, '**' indicates p-values are smaller than 0.01, and '*' indicates p-values are smaller than 0.05. Among all components and interactions, the mapping algorithm contributes the most to the overall variance. Note that more than a quarter of the overall variance belongs to residuals that cannot be explained by the factors considered.

3.3.2.5 Summary of Case Study

We applied the 278 representative RNA-seq pipelines we investigated in Chapter 2, Case Study 4 to the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets and examined whether the choice of various RNA-seq pipeline components would lead to variations in prediction performance of disease outcome. Our results showed that RNA-seq pipeline components jointly affected prediction performance of disease outcome. These joint effects had not previously been reported in any studies. Unlike results presented in Chapter 2, Case Study 4, no single RNA-seq pipeline factors contributed dominantly to the overall variance of the prediction performance. However, many factors such as mapping algorithm, quantification, normalization, and some of their interactions had statistically significant contribution to the overall variance, and normalization usually contributed the most.

3.4 Summary and Key Innovations

In this chapter, I have addressed the second specific aim of this dissertation by designing experiments and implementing statistical modeling and machine learning techniques that facilitate robust knowledge discovery using features extracted from raw -omic data. Knowledge discovery for -omic data can be categorized into two classes—biomarker identification using statistical models and predictive modeling using supervised learning techniques. The chapter started with the introduction of popular statistical modeling and supervised learning frameworks specifically tailored to RNA-seq and ChIP-seq data. These methods were elaborated in the two case studies that focused on two major disease categories—cardiovascular diseases and cancers.

In the first case study, I studied various statistical modeling techniques for both RNA-seq and ChIP-seq datasets that facilitate biomarker identification (i.e., DEGs for the RNA-seq dataset and peaks for the ChIP-seq dataset). Though many methods were publicly available, only a few achieved both statistical and biological significance. Thus, through my studies, I established guidelines for robust DEG detection using RNA-seq and peak calling using ChIP-seq. In the second case study (i.e., the continuation of the SEQC project), we implemented a predictive modeling framework (i.e., the nested cross-validation framework) that produced unbiased, robust prediction performance estimation for several disease endpoints we studied. We found that RNA-seq pipeline components jointly impacted the performance of predictive modeling.

The key innovations of the work in this chapter are listed as follows:

- I conducted the first DEG detection benchmarking study emphasizing both quantitative behavior and biological interpretation of DEGs.
- I performed the largest RNA-seq pipeline investigation so far for two well-known biological datasets and identified key contributing factors for pipeline variability.

CHAPTER 4

INTEGRATIVE ANALYSIS FOR PRECISION MEDICINE

4.1 Introduction

The scope of my dissertation centers on promoting precision medicine via addressing challenges in quality control (Chapter 2), knowledge discovery (Chapter 3), and integrative analysis (to be covered in Chapter 4). In Chapter 3, I introduced several techniques that facilitate knowledge discovery from good quality -omic features. -Omic data provide comprehensive annotations, maps, and catalogs that are beneficial for describing complicated dynamics in the human body. However, each type of -omic data captures only one aspect of molecular dynamics. To understand and model the entire dynamics, a combination of multi-omic data is necessary so as to provide adequate information for inferring and interpreting the true dynamics in cells. The concept of systems biology has been introduced and widely referred since 2000. The idea is to build the system-level understanding of the structure and dynamics in cells rather than inferring functions of a small part of the system using limited information [195]. In Chapter 4, I aim to address the third specific aim by integrating multiple sources of -omic data for improved disease subgroup assignment.

There are many levels of integrative analysis. Following the DIKW hierarchy, the integration can be at the raw data level, information or feature level, knowledge level, or wisdom (i.e., actionable knowledge) level. This chapter focuses on knowledge-level integration aiming to improve the robustness of RNA-seq pipeline recommendation. The work in this chapter is in preparation for submission to *Nature Methods*.

4.2 Knowledge Integration Improving Pipeline Recommendation

4.2.1 Background

This case study is also part of the SEQC project, and it is the follow-up and integration of information described in both Case Study 4 in Chapter 2 (i.e., Section 2.4.4) and Case Study 2 in Chapter 3 (i.e., Section 3.3.2). Please refer to Sections 2.4.4 and 3.3.2 for detailed information about the RNA-seq pipeline study focusing on the benchmark datasets and the cancer datasets, respectively.

The FDA first coordinated multiple sites of SEQC to generate the SEQC-benchmark [148] and SEQC-neuroblastoma datasets [194], and then provided these datasets to SEQC teams to investigate the joint impact of pipeline components on downstream gene expression-based prediction in a two-phase study:

In Phase-1, we developed benchmark metrics—accuracy, precision, reliability, and reproducibility—for assessing a representative set of 278 RNA-seq pipelines (**Figure 46**, blue box) using the SEQC-benchmark dataset.

In Phase-2, we validated the benchmark metrics by quantifying gene expression in the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets, and then determining if the benchmark metrics are informative for inferring downstream prediction of disease outcome (**Figure 46**, pink box).

Our investigation revealed that RNA-seq pipeline components—mapping, quantification, and normalization—jointly impacted the accuracy, precision, reliability, and reproducibility of gene expression, and consequently, affected downstream performance of predicting disease outcome. RNA-seq pipelines that performed well in gene expression estimation also performed well in downstream prediction.

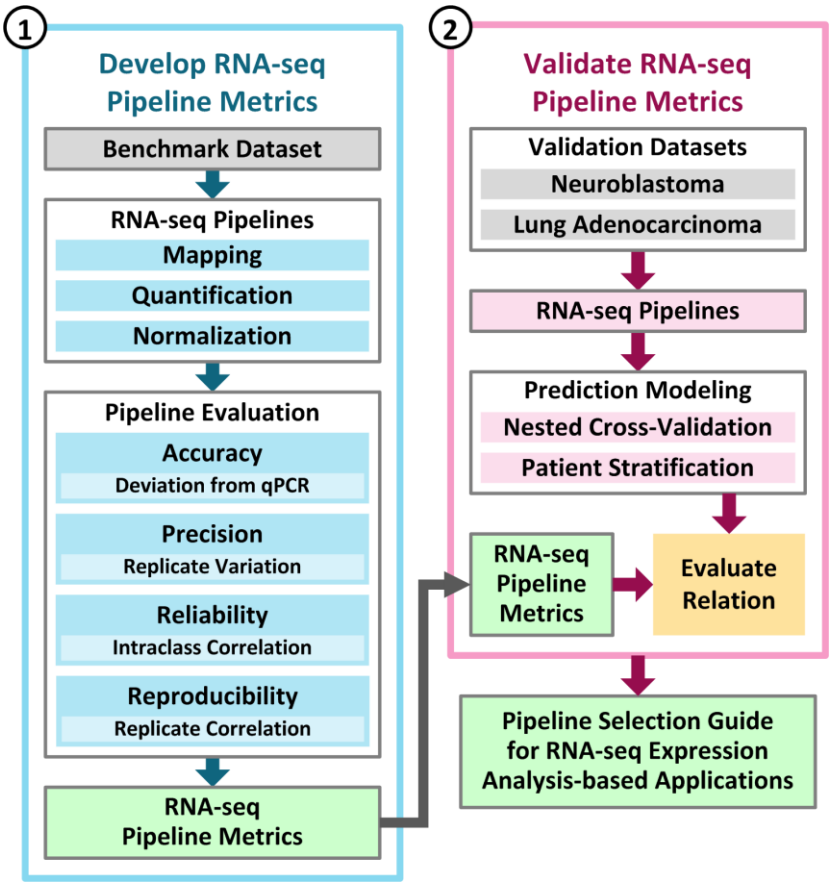


Figure 46: The Workflow of the SEQC Project. The SEQC consortium developed and validated a guideline for selecting RNA-seq pipelines for gene expression-based predictive modeling using the SEQC-benchmark, SEQC-neuroblastoma, and TCGA-lung-adenocarcinoma datasets. Phase-1 of the investigation involved development of metrics that captured the accuracy, precision, reliability, and reproducibility of RNA-seq pipelines (the blue box). Using the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets, Phase-2 of the investigation determined that RNA-seq pipeline metrics can be used to select pipelines that result in better performance in terms of predicting cancer outcome (the pink box).

4.2.2 Experimental Design

The objective of this case study is to show that integrating knowledge derived from various datasets (i.e., the SEQC-benchmark, SEQC-neuroblastoma, and TCGA-lung-adenocarcinoma datasets) can improve the robustness of RNA-seq pipeline recommendations. As illustrated in **Figure 47**, for each dataset, we first ran the 278 or the 156 RNA-seq pipelines, followed by calculating benchmark metrics for the SEQC-benchmark dataset and estimating prediction performance for the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets. Knowledge derived from these datasets is the pipeline ranking based on each individual dataset. Pipeline ranking is dataset-dependent and information-depending (i.e., using benchmark metric performance or prediction performance as the reference). Thus, to provide a robust set of pipelines for RNA-seq expression analysis, it is necessary to integrate knowledge from multiple sources as indicated by the blue box in **Figure 47**.

For the SEQC-benchmark dataset, there are eight benchmark metrics—accuracy, precision, reliability, and reproducibility for both all genes and low-expressing genes. The ranking of each benchmark metric can be encoded by either the percentage (0% represents the best-performing pipeline) or [1, 0, -1] (1: pipeline ranking smaller than 20%; -1: pipeline ranking larger than 80%; 0: otherwise). A pipeline belongs to the good-performing pipelines if either the median of the eight individual percentage ranking is less than 20%, or the sum of [1, 0, -1] ranking is greater than 4. In contrast, a pipeline belongs to the poor-performing pipelines if either the median of the eight individual percentage ranking is greater than 80%, or the sum of [1, 0, -1] ranking is less than -4. We use the same concept to determine the good-performing pipelines and poor-

performing pipelines based on the prediction performance of the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets. The final set of good-performing and poor-performing pipelines are based on the knowledge from all sources.

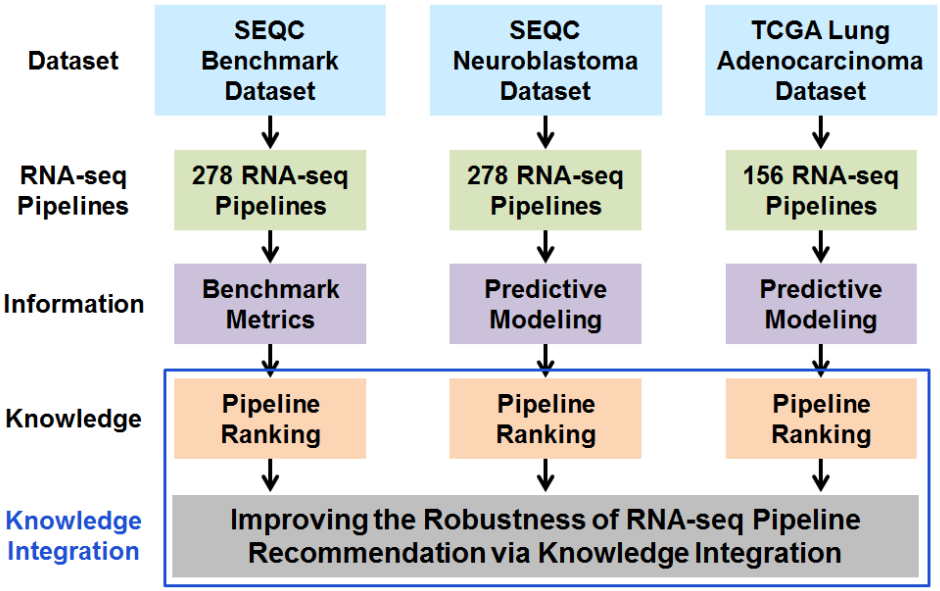


Figure 47: Illustration of Knowledge Integration for Pipeline Recommendation.

Besides identifying a robust set of RNA-seq expression analysis pipelines, we also want to demonstrate that good- and poor-performing RNA-seq pipelines selected based on the benchmark metrics would lead to good and poor prediction performance of disease outcome, respectively (**Figure 48**).

To achieve this objective, we first rank RNA-seq pipelines base on the median rank of the four benchmark metrics (i.e., accuracy, precision, reliability, and reproducibility). We then evaluate the utility of the benchmark metrics by examining whether good-performing and poor-performing pipelines identified based on the benchmark metrics were informative for inferring the performance of gene-expression-

based prediction of disease outcome and statistical significance of patient stratification for all clinical endpoints (i.e., the SEQC-neuroblastoma EFS and OS endpoints and the TCGA-lung-adenocarcinoma survival endpoint).

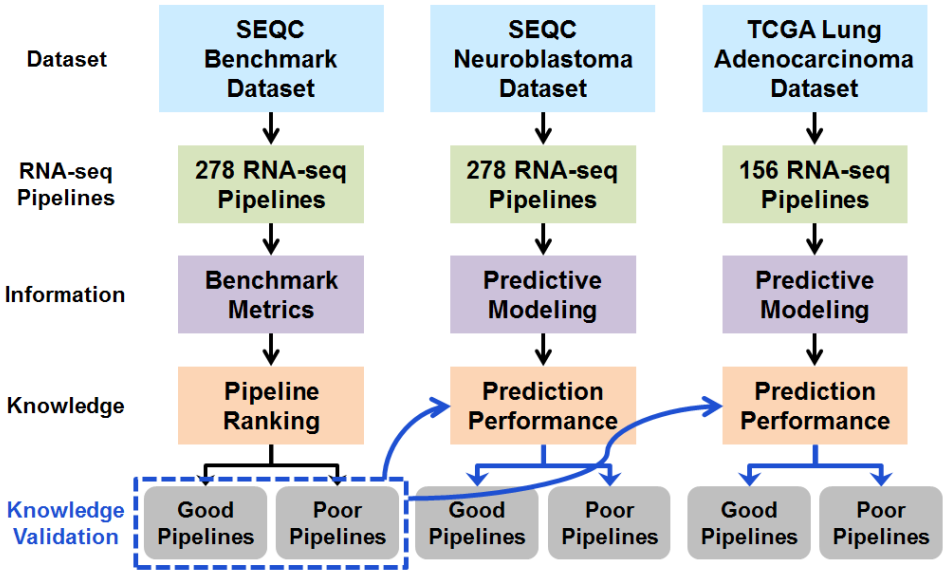


Figure 48: Illustration of Knowledge Validation for Pipeline Selection.

First, for the 278 representative RNA-seq pipelines applied to the SEQC-benchmark dataset, we compute the median rank using a subset of the benchmark metrics as the final performance indicator for each pipeline. In total, we have 8 metrics (4 benchmark metrics [accuracy, precision, reliability, reproducibility] × 2 gene sets [2,044 low-expressing genes, 10,222 all genes]), and we investigate 45 subsets (3×15) of the 8 metrics using the following criteria:

(1) Fifteen combinations of the four benchmark metrics with at least one in a subset—one combination with all four benchmark metrics, four combinations

with three out of the four benchmark metrics, six combinations with two out of the four benchmark metrics, and four combinations with one metric.

(2) Three subsets formed by metrics derived from all genes, those derived from low-expressing genes, or a combination of both.

Second, for each of the 278 representative RNA-seq pipelines (156 for the TCGA-lung-adenocarcinoma survival endpoint), we calculate nested cross-validation AUC and MCC, resulting in 834 (468 for the TCGA-lung-adenocarcinoma survival endpoint) AUC and MCC values for each clinical endpoint (i.e., 278 pipelines × 3 classifiers, or 156 pipelines × 3 classifiers). We also model survival functions based on the predicted labels of each sample using Kaplan-Meier analysis for each pipeline and each classifier. We then use the two-tailed log-rank test to determine if estimated survival curves between the two predicted patient groups were statistically different. For each RNA-seq pipeline, we summarize the performance of gene-expression-based prediction of disease outcome using both the average AUC and MCC across classifiers and the success rate of patient stratification (i.e., statistically significant separation of two Kaplan-Meier curves) across all iterations and classifiers.

Finally, we identified the top 20% good-performing pipelines and the bottom 20% poor-performing pipelines based on the median rank of a subset of the four benchmark metrics. The corresponding prediction performance (i.e., AUC and MCC) of the good-performing pipelines was tested against that of the poor-performing pipelines using the one-sided Wilcoxon rank-sum test with the null hypothesis that the median of the former group was not larger than that of the latter group.

4.2.3 Results and Discussion

For robust RNA-seq pipeline recommendation, the good-performing and poor-performing RNA-seq pipelines are summarized in **Figure 49**. Key factors that contributed to good and poor performance were also listed.

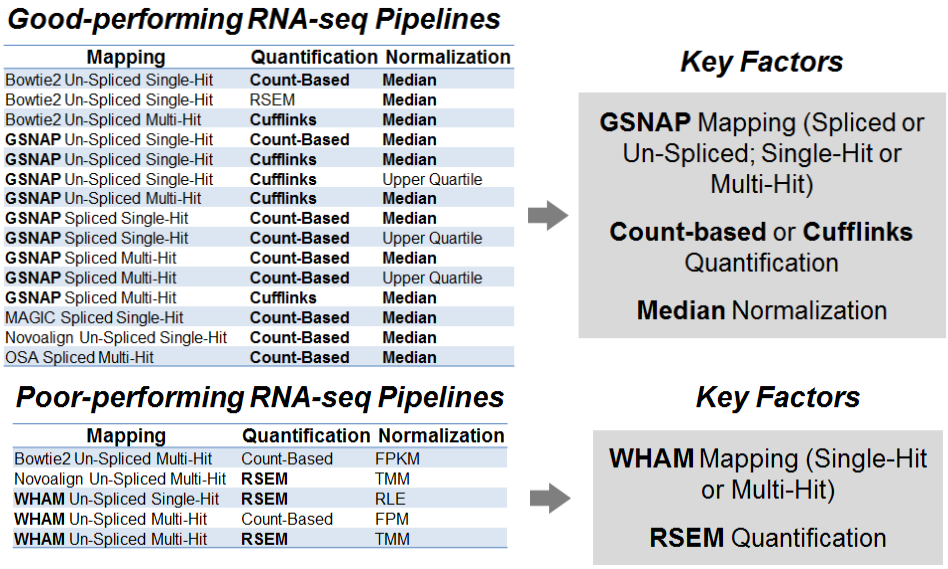


Figure 49: A Robust Set of Good- and Poor-Performing RNA-seq Pipelines.

For knowledge validation, we ranked the 278 RNA-seq pipelines base on the median rank of a combination of the four benchmark metrics. We then investigated if good-performing and poor-performing pipelines identified based on the benchmark metrics were informative for inferring the performance of gene-expression-based prediction of disease outcome.

For all endpoints, median prediction performance (i.e., AUC and MCC) of good-performing pipelines was statistically significantly ($p<0.05$) larger than that of poor-performing pipelines (**Figure 50**) based on the one-sided Wilcoxon rank-sum test. In

addition, good-performing pipelines (e.g., the [GSNAP un-spliced single-hit + Cufflinks + median] pipeline) tended to result in higher success rates of patient stratification than poor-performing pipelines (e.g., the [BWA + RSEM + RLE] pipeline). **Figure 51** demonstrates Kaplan-Meier estimated survival functions for high-risk and low-risk patients for all endpoints. Good-performing pipelines tended to achieve statistically significant separation ($p<0.05$) of the two patient groups (**Figure 51**, panels a – c), while poor-performing pipelines were more likely to fail (**Figure 51**, panels d – f).

In this case study, by integrating knowledge (i.e., good-performing and poor-performing RNA-seq pipelines) from the three datasets, we were able to show that pipelines that performed well in the SEQC-benchmark dataset would also performed well in the SEQC-neuroblastoma and TCGA-lung-adenocarcinoma datasets. With the results demonstrated here, it becomes valid to recommend a robust set of RNA-seq pipelines that will perform well or poorly for multiple RNA-seq datasets with very different data-dependent characteristics.

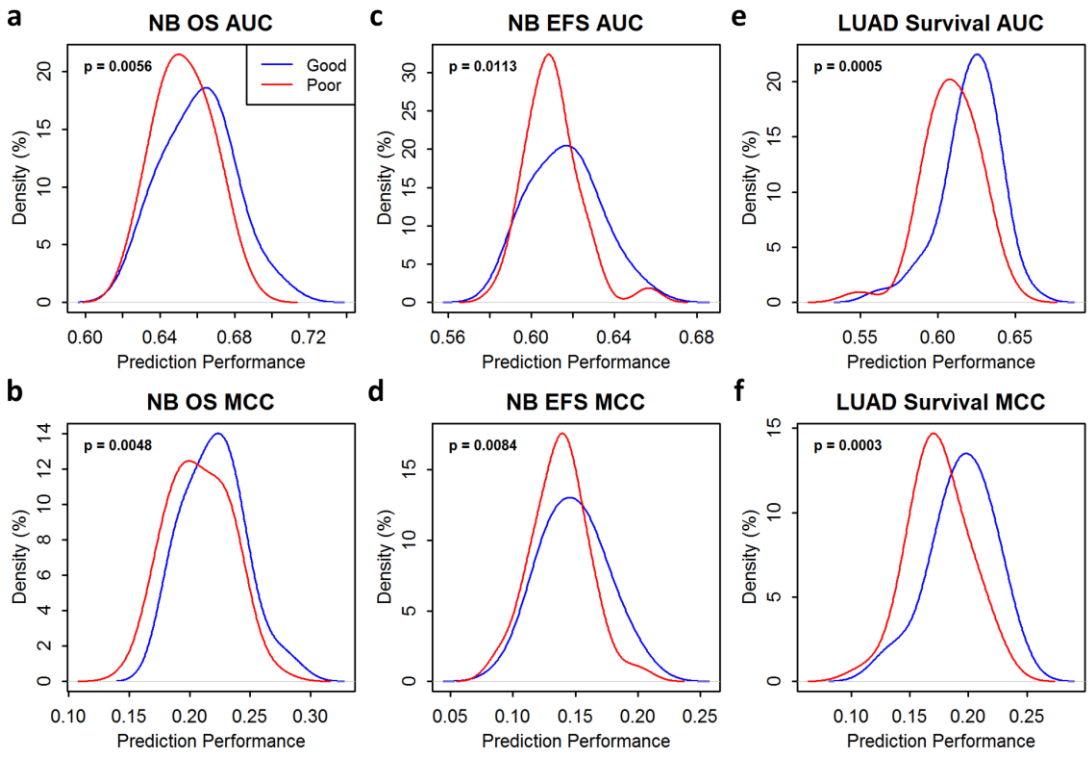


Figure 50: Knowledge Validation for Pipeline Selection—Prediction Performance. RNA-seq pipelines selected based on benchmark metrics (i.e., accuracy, precision, reliability, and reproducibility) were informative for inferring the performance of gene-expression-based prediction of disease outcome—(a) prediction performance measured by the area under the receiver operating characteristic curve (AUROC, or AUC) for the overall survival (OS) endpoint of the SEQC-neuroblastoma (NB) dataset; (b) prediction performance measured by the Matthews correlation coefficient (MCC) for the OS endpoint of the SEQC-NB dataset; (c) prediction performance measured by the AUC for the event-free survival (EFS) endpoint of the SEQC-NB dataset; (d) prediction performance measured by the MCC for the EFS endpoint of the SEQC-NB dataset; (e) prediction performance measured by the AUC for the survival endpoint of the TCGA-lung-adenocarcinoma (LUAD) dataset; and (f) prediction performance measured by the MCC for the survival endpoint of the TCGA-LUAD dataset. The blue line in each panel shows the probability density of the prediction performance of good-performing RNA-seq pipelines selected based on benchmark metrics; and the red line demonstrates that of poor-performing pipelines selected based on the same. Statistical significance (i.e., p-values) was determined using the one-sided Wilcoxon rank-sum test. All panels show statistically significant difference between the two groups (i.e., prediction performance of good-performing pipelines vs. that of poor-performing pipelines). The good-performing and poor-performing pipelines were determined based on the average rank of each RNA-seq pipeline over all benchmark metrics of low-expressing genes.

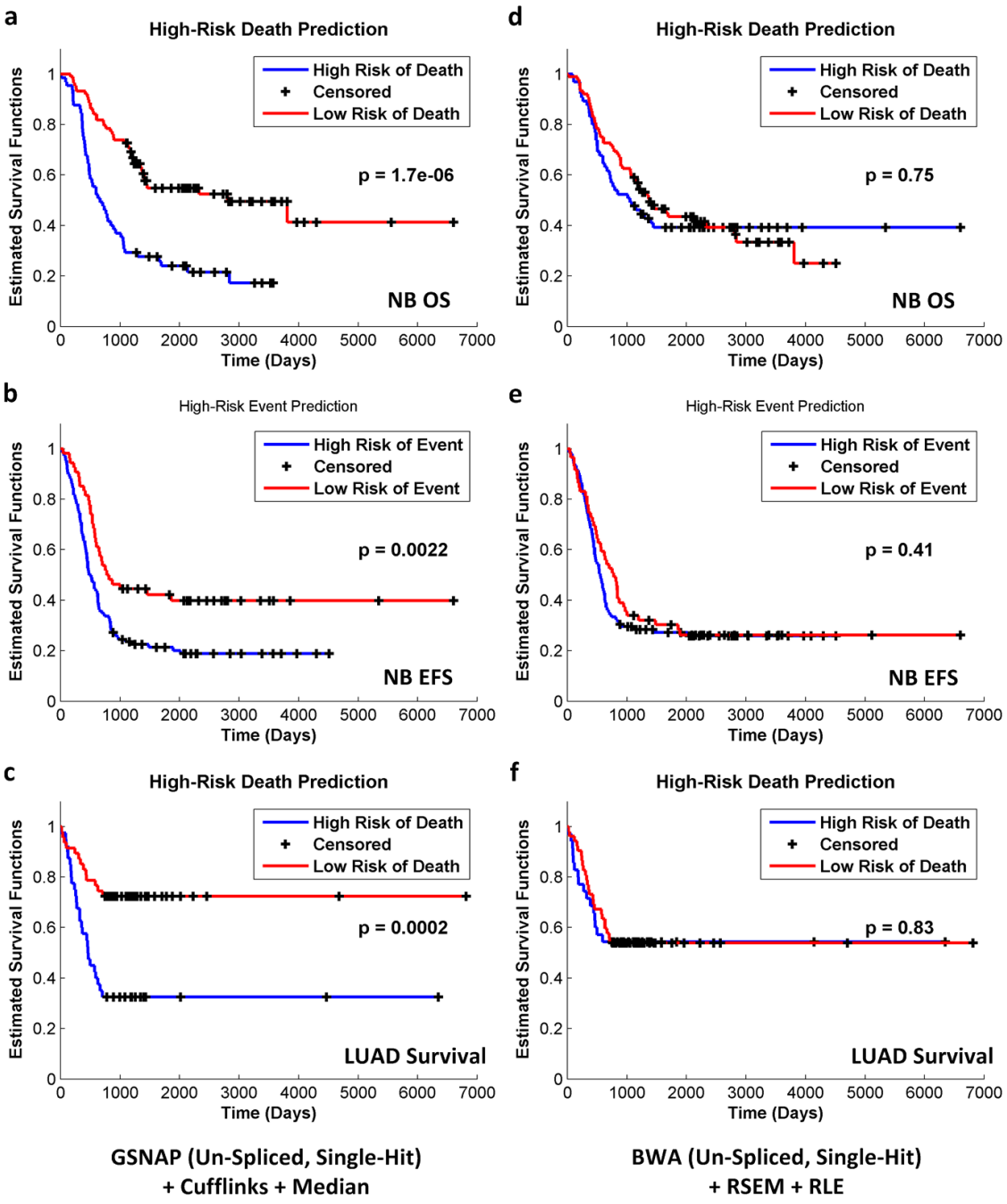


Figure 51: Knowledge Validation for Pipeline Selection—Kaplan-Meier Analysis. The RNA-seq pipeline selection guide was validated by assessing the ability of pipelines to stratify patients based on Kaplan-Meier survival analysis. For each pipeline, patients were grouped by predictive labels (i.e., high risks vs. low risk), and two Kaplan-Meier curves were plotted. The two-tailed log-rank test was used to determine the statistical significance of the difference between the two curves. For good-performing pipelines selected based on benchmark metrics, the success rates of patient stratification (i.e., predictive labels led to statistically significant separation of Kaplan-Meier curves) were higher. For example, the success rates of the [GSNAP (un-spliced, single-hit) + Cufflinks

+ Median] pipeline were 93%, 70%, and 67% for the SEQC-NB OS, SEQC-NB-EFS, and TCGA-LUAD-Survival endpoints, respectively. Panels (a) to (c) demonstrate the most statistically significant separation of the two Kaplan-Meier curves for each endpoint. In contrast, poor-performing pipelines led to lower success rates of patient stratification. For instance, the success rates of the [BWA (un-spliced, single-hit) + RSEM + RLE] pipeline were 33%, 30%, and 33% for the SEQC-NB OS, SEQC-NB-EFS, and TCGA-LUAD-Survival endpoints, respectively. Panels (d) to (f) demonstrate the least statistically significant separation for each endpoint.

4.4 Summary and Key Innovations

In this chapter, I have addressed the third specific aim of this dissertation by designing experiments that integrate knowledge from multiple datasets of the same type. The knowledge refers to pipeline rankings based on different datasets. Our integrative analysis provided a foundation for more robust, reliable RNA-seq pipeline recommendations.

The key innovations of the work in this chapter are listed as follows:

- I designed a knowledge integration workflow incorporating RNA-seq pipeline rankings based on different RNA-seq datasets that improves the robustness of RNA-seq pipeline recommendations.

CHAPTER 5

CONCLUSION

The concrete goals of this dissertation were to investigate and develop integrative bioinformatics approaches for extracting and discovering robust molecular knowledge for realizing future precision medicine. The specific technical achievements of this dissertation corresponding to the three research objectives are as follows:

1. Investigated the impact of pipeline choice—component-wise and pipeline-wise—on the quality of gene/transcript expression estimates for RNA-seq data; and established quality control guidelines for RNA-seq expression analysis pipelines
2. Identified significant, predictive biomarkers for various clinical settings using statistical modeling and predictive modeling techniques; and established knowledge discovery guidelines for DEG detection
3. Designed and implemented a workflow integrating pipeline rankings based on multiple RNA-seq datasets

5.1 Concrete Innovation Deliverables

The key innovations of this dissertation, as noted at the closing of each chapter, are summarized as follows:

- (Chapter 2) I designed a comprehensive list of evaluation metrics that capture the performance of RNA-seq expression analysis pipeline.
- (Chapter 2) I conducted the first investigation on genome annotation and proposed a novel, informative annotation complexity measure.

- (Chapter 2) I performed quantification pipeline investigation (among the first batch) and identified key factors for achieving accurate expression estimates.
- (Chapter 2) I accomplished the simulation-based investigation on expression normalization (among the first batch).
- (Chapter 2) I performed the largest investigation of RNA-seq expression analysis pipeline so far using well-designed benchmark datasets provided by FDA.
- (Chapter 3) I conducted the first DEG detection benchmarking study emphasizing both quantitative behavior and biological interpretation of DEGs.
- (Chapter 3) I performed the largest RNA-seq pipeline investigation so far for two well-known biological datasets and identified key contributing factors for pipeline variability.
- (Chapter 4) I designed a knowledge integration workflow incorporating RNA-seq pipeline rankings based on different RNA-seq datasets that improves the robustness of RNA-seq pipeline recommendations.

5.2 Directions for Future Research and Concluding Remarks

5.2.1 -Omic Data Integration

One notable effort that integrates multi-omic data for the improved understanding of cancer mechanisms is The Cancer Genome Atlas (TCGA) [196]. TCGA hosts public datasets of 27 cancer types with more than 11,000 patient cases. Each patient is annotated with clinical data (i.e. demographic, diagnostic, and survival data) and multimodal -omic data (i.e., genomic, transcriptomic, epigenomic, and proteomic).

We use head and neck squamous cell carcinoma (HNSCC) as an example to illustrate the integrative multi-omic study for precision medicine [197]. In 2014, a pan-cancer study with twelve cancer types using multi-omic TCGA data was performed [198]. Among 3,527 samples in total, 305 were HNSCC. Six different data types (i.e. DNA copy number, methylation, mutation, and expression of mRNAs, miRNAs, and proteins) were analyzed both separately and integratively. By using clustering-based methods, pathway activities (inferred from gene expression and copy number data) have shown common copy number variations, mutation frequency patterns, and survival patterns between HNSCC and lung squamous cell carcinomas or some bladder cancers. Such integrative pan-cancer analysis provides more precise subtyping across multiple cancers sharing common molecular-level processes underlying cancer development. This new subtyping system reflects the essence of precision medicine.

TCGA Research Network has published more than 30 articles describing multi-omic investigation on numerous cancer types, and identified more precise, clinically relevant subtyping for multiple cancers [199-201]

As illustrated by the TCGA case study, integrative multi-omic data analysis is of growing importance because it provides holistic view of molecular fingerprints for each patient's condition. Recent research has shown positive impact of knowledge and insight obtained from integrative analysis of genomic and transcriptomic [202], transcriptomic and proteomic [203], and multiple -omic data types [111, 198] on disease diagnosis, prognosis, and treatment. The next important direction is the development of guidelines (or best practices) for -omic data integration and interpretation that will in turn enable better prediction of bio-system behavior, and safer and more effective therapeutics.

5.2.2 -Omic Data in EHR

In a clinical setting, healthcare providers use electronic medical record (EMR) for clinical decision support. Thus, it is important to incorporate -omic data and knowledge into EMR. The Electronic Medical Records and Genomics (eMERGE) Network consortium aims to identify causal genomic variants (mostly SNPs) for EMR-based phenotypes and to integrate identified genotype-phenotype associations into the EMR system [204]. One crucial challenge is on how to store variants present in an individual or even in family members in the EMR [205]. The consortium has proposed several recommendations on augmenting the current EMR structure: (1) it should store various genomic variants, such as SNPs, indels, and CNVs, in a discrete computable format; (2) it needs to satisfy interoperability to reduce the burden in data transfer and update within and between healthcare facilities; (3) it has to support rule-based decision support engines; and (4) it must contain abundant visualization elements for easier interpretation [206]. Another big challenge is that each individual typically has millions of variants. The consortium has proposed one potential solution that stores only known pathological variants in the EMR system. However, because the set of known pathological variants may change over time, this approach may lead to the inclusion of false positive and the exclusion of false negative variants. Thus, an alternative solution is to archive raw data in separate repositories easily accessible when necessary [207].

EMR is only for local clinic and hospital, while EHR contains and shares medical records among all participant clinics and hospitals [208]. Thus, interoperability is critical in using big data for precision medicine. Recently, the Health Level Seven International (HL7) proposed the Fast Healthcare Interoperability Resources (FHIR) standard that

addresses this important issue. On clinical genomics, several new FHIR resources and extension definitions are designed for variant data [209]. With such the standardized data exchange protocol, clinicians can utilize genomic information with other existing EHR data to determine the most effective treatment for each patient, which is a paradigm shift towards precision medicine.

5.2.5 Concluding Remarks

In this dissertation, I have addressed challenges in three fundamental building blocks, that is, quality control, knowledge discovery, and integrative analysis, for precision medicine, with the emphasis on analytics and models for NGS data. In the preceding sections, I have also discussed several potential directions building upon this work. Overall, this dissertation contributes to the research space by laying a foundation for future precision medicine.

APPENDIX A

RELEVANT PUBLICATIONS

The work presented in this dissertation is a culmination of several years of research that resulted in the following peer reviewed journal articles, book chapters, and conference proceedings.

In Preparation/Submitted

(JP-1) **Wu PY**, Phan JH, Maher KO, Mahle WT, Wang DM. Integration of Genomic and Proteomic Data for Clinical Outcome Prediction after Neonatal Cardiac Surgery. *Circulation: Cardiovascular Genetics*. In Preparation.

(JP-2) Phan JH, **Wu PY**, MAQC-III Consortium. Impact of RNA-seq data analysis algorithms on gene expression estimation and downstream prediction. *Nature Methods*. In Preparation.

Journal Articles

(J-1) **Wu PY**, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. Advanced big data analytics for -omic data and electronic health records: toward precision medicine. *IEEE Transactions on Biomedical Engineering*. In Press.

(J-2) Zarkogianni K, Litsa E, Mitsis K, **Wu PY**, Kaddi CD, Cheng CW, Wang MD, Nikita KS. A review of emerging technologies for the management of diabetes mellitus. *IEEE Transactions on Biomedical Engineering*. 2015 Aug;62(12):2735-2749.

(J-3) Zhang W, Yu Y, MAQC-III Consortium. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*. 2015 Jun;16(133).

- **Wu PY** is a contributing author of this paper.

(J-4) **Wu PY**, Chandramohan R, Phan JH, Mahle WT, Maher K, Gaynor JW, Wang MD. Cardiovascular transcriptomics and epigenomics using next-generation sequencing—challenges, progress, and opportunities. *Circulation: Cardiovascular Genetics*. 2014 Oct;7(5):701-710.

(J-5) SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*. 2014 Aug;32(9):903-914.

- **Wu PY** is a contributing author of this paper.

(J-6) Li S, Labaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan JH, **Wu PY**, Wang MD, Wang C, Thierry-Mieg D, Thierry-Mieg J, Kreil DP, Mason CE. Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nature Biotechnology*. 2014 Aug;32(9):888-895.

(J-7) **Wu PY**, Phan JH, Wang MD. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC Bioinformatics*. 2013 Nov;14(Suppl 11):S8.

(J-8) Zhang H, **Wu PY**, Ma M, Ye Y, Hao Y, Yang J, Yin S, Sun C, Phan JH, Wang MD, Xi JJ. An integrative approach for the large-scale identification of human genome kinases regulating cancer metastasis. Nanome*dicine: Nanotechnology, Biology and Medicine*. 2013 Aug;9(6):732-736.

Conference Proceedings

(C-1) **Wu PY** and Wang MD. The selection of quantification pipelines for Illumina RNA-seq data using a subsampling approach. *Proceedings of the 2016 International Conference on Biomedical and Health Informatics*. Las Vegas, NV. 2016 Feb;78-81.

(C-2) Phan JH, Hoffman R, Kothari S, **Wu PY**, Wang MD. Integration of multi-modal biomedical data to predict cancer grade and patient survival. *Proceedings of the 2016 International Conference on Biomedical and Health Informatics*. Las Vegas, NV. 2016 Feb;577-580.

(C-3) Tong L, Yang C, **Wu PY**, Wang MD. Evaluating the impact of sequencing error correction for RNA-seq data with ERCC RNA spike-in controls. *Proceedings of the 2016 International Conference on Biomedical and Health Informatics*. Las Vegas, NV. 2016 Feb;74-77.

(C-4) **Wu PY**, Phan JH, Wang MD. Prediction of cardiac ICU length of stay after infant cardiac surgery using exome sequencing data. *NIH-IEEE 2015 Strategic Conference on Healthcare Innovations and Point-of-Care Technologies for Precision Medicine*. Bethesda, MD. 2015 Nov.

(C-5) Yang C, **Wu PY**, Tong L, Phan JH, Wang MD. The impact of RNA-seq aligners on gene expression estimation. *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*. Atlanta, GA. 2015 Sept;462-471.

(C-6) Yang C, **Wu PY**, Phan JH, Wang MD. The impact of RNA-seq alignment pipeline on detection of differentially expressed genes. *Proceedings of the 2014 IEEE*

International Workshop on Genomic Signal Processing and Statistics. Atlanta, GA. 2014 Dec;1376-1379.

(C-7) **Wu PY**, Phan JH, Wang MD. An approach for assessing RNA-seq quantification algorithms in replication studies. *Proceedings of the 2013 IEEE International Workshop on Genomic Signal Processing and Statistics*. Houston, TX. 2013 Nov;15-18.

(C-8) Cheng CW, Martin GS, **Wu PY**, Wang MD. PHARM—association rule mining for predictive health. *IFMBE Proceedings of the International Conference on Health Informatics*. Vilamoura, Portugal. 2013 Nov;114-117.

(C-9) Chandramohan R, **Wu PY**, Phan JH, Wang MD. Systematic assessment of RNA-seq quantification tools using simulated sequence data. *Proceedings of the 4th ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. Washington, DC. 2013 Sept.623-632.

(C-10) Chandramohan R, **Wu PY**, Phan JH, Wang MD. Benchmarking RNA-seq quantification tools. *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Osaka, Japan. 2013 Jul;647-650.

(C-11) Phan JH, **Wu PY**, Wang MD. Improving the flexibility of RNA-seq data analysis pipelines. *Proceedings of the 2012 IEEE International Workshop on Genomic Signal Processing and Statistics*. Washington, DC. 2012 Dec;70-73.

(C-12) **Wu PY**, Phan JH, Wang MD. The effect of human genome annotation complexity on RNA-seq gene expression quantification. *Proceedings of the 2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. Philadelphia, PA. 2012 Oct;712-717.

(C-13) **Wu PY**, Phan JH, Zhou F, Wang MD. Evaluation of normalization methods for RNA-seq gene expression estimation. *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. Atlanta, GA. 2011 Nov;50-57.

(C-14) **Wu PY**, Phan JH, Wang MD. Exploring the feasibility of next-generation sequencing and microarray data meta-analysis. *Proceedings of the 33th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Boston, MA. 2011 Aug;7618-7621.

(C-15) Srimani JK, **Wu PY**, Phan JH, Wang MD. A distributed system for fast alignment of next-generation sequencing data. *Proceedings of the 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. Hong Kong. 2010 Dec;579-584.

APPENDIX B

SUPPLEMENTARY NOTES FOR THE SEQC PROJECT

Filtering the qPCR Benchmark Dataset to Produce a Reference Set of Genes

Due to variability in qPCR measurements as well as disagreements among qPCR platforms [148], we filtered the qPCR dataset to retain only genes that exhibited "reliable" behavior. We then used these genes to evaluate the RNA-seq pipeline metrics. Filtering of the qPCR gene set is summarized in **Figure 52**.

Starting with the initial set of 20,801 genes assayed with PrimePCR, we filtered these genes to retain only genes that were quantified as non-zero (i.e., detected) and with Ct (cycle threshold) values ≤ 35 (35 indicates detection of only a single molecule). Filtering PrimePCR data resulted in 14,014 genes that also matched with the AceView transcriptome used for mapping the SEQC-benchmark datasets with RNA-seq pipelines.

Subsequently, we filtered the 14,014 qPCR genes to retain only those that exhibited the correct titration order (TO) and expected mixing ratio (EMR). Details of this process are in the following section titled "Filtering qPCR Genes by Titration Order and Expected Mixing Ratios." 12,610 genes were retained after this process.

Lastly, since some benchmark metrics such as accuracy and precision are sensitive to zero- or very low-expressing genes, we further selected genes that were expressed as non-zero in all replicates of all samples of all sequencing sites and all 278 RNA-seq pipelines. The final reference set contains only 10,222 qPCR genes (referred to as "all genes") that were used to compute all four benchmark metrics for RNA-seq pipelines.

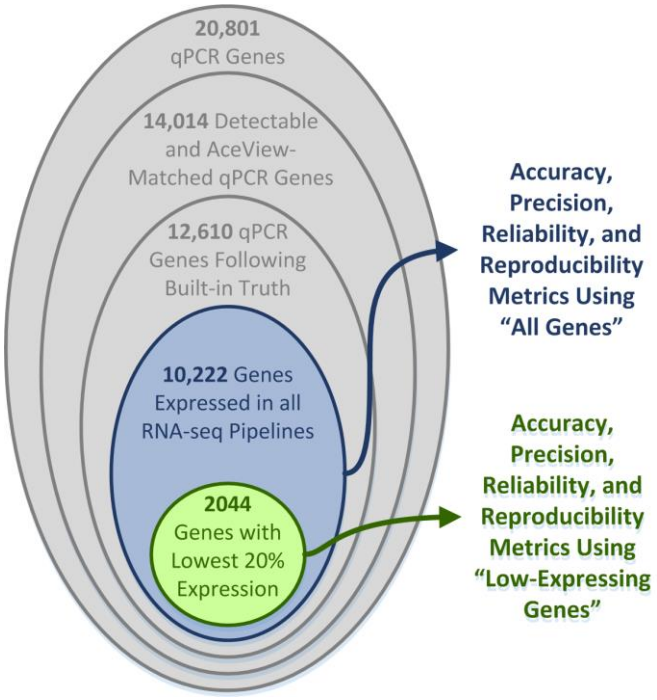


Figure 52: Filtering Benchmark qPCR Genes. The 20,801 qPCR genes are first filtered to retain only 14,014 genes that are detectable and that match genes in the AceView transcriptome. Subsequently, genes are filtered to retain only the 12,610 genes that exhibit good titration order and expected mixing ratios. Finally, 10,222 genes (denoted as "All Genes") that expressed in all replicates of all samples of all 278 RNA-seq pipelines are retained for calculating evaluation metrics (i.e., accuracy, precision, reproducibility, and reliability). Among 10,222 genes, the 20% lowest-expressing genes (i.e., 2,044 genes; denoted as "Low-Expressing Genes") are selected for calculating another set of metrics.

We also identified a set of low-expressing genes in the 10,222 genes based on the average qPCR expression of samples A, B, C, and D. The lowest 20% of the 10,222 genes (i.e., 2,044 genes, referred to as "low-expressing genes") were also used to compute the same set of benchmark metrics for RNA-seq pipelines. This design enabled us to investigate the capability of RNA-seq pipelines in estimating low-expressing gene expression.

Filtering qPCR Genes by Titration Order and Expected Mixing Ratios

The SEQC-benchmark datasets (RNA-seq and qPCR) have unique properties that enable assessment of quantification reliability. After identifying detectable and AceView-matched qPCR genes, we used two metrics (TO and EMR) to further filter the qPCR data, leaving only "reliable" qPCR genes. The TO and EMR metrics capture unique mixing properties of the data, that is,

$$C = \frac{3}{4}A + \frac{1}{4}B \quad \text{and} \quad D = \frac{1}{4}A + \frac{3}{4}B \ . \tag{22}$$

Because of this property, all genes are expected to be expressed in one of the following orders, depending on the relative expression of samples A and B:

$$A \geq C \geq D \geq B \quad \text{or} \quad A \leq C \leq D \leq B \ . \tag{23}$$

The TO metric determines if genes are expressed in the correct order. The expression value of a qPCR gene is defined as $y_{s,n,k}$ where $s \in \{A, B, C, D\}$ indicates the sample, $n = 1 \ldots N$ indicates the replicate, and $k = 1 \ldots K$ indicates the gene (for the PrimePCR set, N = 1; K = 10,222). For a qPCR dataset with multiple replicates, given that the mean expression value for gene k, and sample s over all replicates as

$$\bar{y}_{s,\cdot,k} = \frac{1}{N}\sum_{n=1}^{N} y_{s,n,k} \ , \tag{24}$$

the set of all qPCR genes that follow the correct TO is

$$K_{TO} = \left\{k \middle| \left(\bar{y}_{A,\cdot,k} \geq \bar{y}_{C,\cdot,k} \geq \bar{y}_{D,\cdot,k} \geq \bar{y}_{B,\cdot,k}\right) \vee \left(\bar{y}_{A,\cdot,k} \leq \bar{y}_{C,\cdot,k} \leq \bar{y}_{D,\cdot,k} \leq \bar{y}_{B,\cdot,k}\right)\right\} . \tag{25}$$

For a single replicate qPCR dataset (e.g., the PrimePCR dataset we analyzed), inherent variability of a single qPCR measurement may result in some false negative genes that follow the correct TO but fail to be identified. From the literature [210-213], the typical CoV for replicate qPCR measurements is around 15%, so we used this number to increase the margin for determining whether a gene follows the correct TO.

Mathematically, we calculated the width of one standard deviation of each qPCR measurement and used it as the margin. The revised equations for K_{TO} are as follow:

$$a = 1.15,\ b = 0.85, \tag{26}$$

$$K_{TO,A\geq B} = \left\{k \middle| \left(a \cdot \bar{y}_{A,k} \geq b \cdot \bar{y}_{C,k}\right) \wedge \left(a \cdot \bar{y}_{C,k} \geq b \cdot \bar{y}_{D,k}\right) \wedge \left(a \cdot \bar{y}_{D,k} \geq b \cdot \bar{y}_{B,k}\right)\right\},$$

$$K_{TO,A\leq B} = \left\{k \middle| \left(b \cdot \bar{y}_{A,k} \leq a \cdot \bar{y}_{C,k}\right) \wedge \left(b \cdot \bar{y}_{C,k} \leq a \cdot \bar{y}_{D,k}\right) \wedge \left(b \cdot \bar{y}_{D,k} \leq a \cdot \bar{y}_{B,k}\right)\right\},$$

$$K_{TO} = K_{TO,A\geq B} \cup K_{TO,A\leq B}\ .$$

Besides the TO, samples should additionally exhibit a specific mixing ratio. Given that the ratio between samples A and B as

$$R_{A,B} = \frac{A}{B}, \tag{27}$$

the EMR between samples C and D is

$$EMR_{C,D} = \frac{3z \cdot R_{A,B}+1}{z \cdot R_{A,B}+3} \cdot \frac{z+3}{3z+1}, \tag{28}$$

where $z = \frac{mRNA\ Concentration\ in\ A}{mRNA\ Concentration\ in\ B} = 1.43$, a correction factor for the difference in mRNA concentration between samples A and B [148].

The EMR metric examines if $EMR_{C,D}$ of a gene is close enough to observed $R_{C,D} = \frac{C}{D}$ of the same gene. As described earlier, the PrimePCR dataset contains only a single measurement for each sample, and thus, wider margins are needed for EMR metric calculation. Using the same technique, we calculated the width of one standard deviation of each ratio

$$R_{A,B} \in \left[b \cdot R_{A,B},\ a \cdot R_{A,B}\right] \equiv \left[R_{A,B}^{Lower},\ R_{A,B}^{Upper}\right], \tag{29}$$

$$R_{C,D} \in \left[b \cdot R_{C,D},\ a \cdot R_{C,D}\right] \equiv \left[R_{C,D}^{Lower},\ R_{C,D}^{Upper}\right],$$

$$EMR_{C,D} \in \left[b \cdot EMR_{C,D},\ a \cdot EMR_{C,D}\right] \equiv \left[EMR_{C,D}^{Lower},\ EMR_{C,D}^{Upper}\right]),$$

and finally determines a set of genes that satisfies the EMR criterion as follows:

$$K_{EMR} = \left\{k \middle| \left(R_{C,D}^{Lower} \leq EMR_{C,D}^{Upper}|_{k,R_{C,D} \geq EMR_{C,D}}\right) \vee \left(R_{C,D}^{Upper} \geq EMR_{C,D}^{Lower}|_{k,R_{C,D} \leq EMR_{C,D}}\right)\right\}. \quad \textbf{(30)}$$

Regression Analysis

We investigated the relationship between alignment profiles or gene expression distribution characteristics and benchmark metrics. The alignment profiles included the total number of mapped fragments, the total number of reads spanning the intronic region, the total number of reads with insertions or deletions, the total number of perfectly matched reads, the total number of reads with at most one mismatch, and the number of mismatches per mapped read. Each alignment algorithm was represented by the average statistics over 2 sequencing sites, 4 samples, 4 replicate libraries, and 2 lanes. Using the "MASS" package in R, we adopted the M-estimation with Huber weighting approach to fit robust linear regression models between a dependent variable (benchmark metric performance) and an explanatory variable (an alignment profile). The M-estimation with Huber weighting approach is a regression method that is robust in the presence of outliers. The gene expression distribution characteristics included the lower quartile, median, upper quartile, maximum, interquartile range, standard deviation, skewness, kurtosis, and entropy of a gene expression distribution. We used the same M-estimation with Huber weighting approach to fit robust linear regression models, and then reported the residual standard error for each model.

Analysis of Variance for the SEQC Project

We used analysis of variance (ANOVA) to determine if each RNA-seq pipeline factor significantly contributes to the variance of each of the four benchmark metrics (i.e., accuracy, precision, reliability, and reproducibility) as well as to the variance of

prediction performance (i.e., AUC and MCC). For each of the four benchmark metrics, we used a linear model (R function "lm") to fit the data from all 278 pipelines using the metric as the dependent variable and the RNA-seq pipeline factors as independent categorical variables. We considered the following factors as independent categorical variables—mapping algorithm, mapping strategy (i.e., spliced vs. un-spliced), mapping reporting (i.e., single-hit vs. multi-hit), quantification algorithm, and normalization algorithm. We included all factors and their two-way interactions in the linear model. For each of the prediction endpoints, we applied the same technique to fit the data from all 278 pipelines using average AUC or MCC as the dependent variable and the same set of RNA-seq pipeline factors as independent categorical variables. We then conducted the ANOVA on the linear model (R function "anova"). ANOVA calculates a "sum of squares" (i.e., variance) attributed to each factor or interaction and uses an F-test to determine if the variance is statistically significant. We calculated the percent that each factor or interaction contributes to the total variance by calculating the ratio of "sum of squares" for each factor to the total sum of squares.

REFERENCES

[1] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, "Bioinformatics challenges for personalized medicine," *Bioinformatics*, vol. 27, pp. 1741-8, Jul 1 2011.

[2] G. S. Ginsburg and H. F. Willard, "Genomic and personalized medicine: foundations and applications," *Transl Res*, vol. 154, pp. 277-87, Dec 2009.

[3] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," *Nat Rev Clin Oncol*, vol. 8, pp. 184-7, Mar 2011.

[4] S. Desmond-Hellmann, C. L. Sawyers, D. R. Cox, and C. Fraser-Liggett, in *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*, ed Washington (DC): The National Academies Press, 2011.

[5] A. Katsnelson, "Momentum grows to make 'personalized' medicine more 'precise'," *Nature Medicine*, vol. 19, pp. 249-249, Mar 2013.

[6] T. W. House. (2015). *Remarks by the President on Precision Medicine*. Available: https://www.whitehouse.gov/the-press-office/2015/01/30/remarks-president-precision-medicine

[7] R. Mirnezami, J. Nicholson, and A. Darzi, "Preparing for precision medicine," *N Engl J Med*, vol. 366, pp. 489-91, Feb 9 2012.

[8] T. W. House. (2015). *FACT SHEET: President Obama's Precision Medicine Initiative*. Available: https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

[9] C. G. Chute, M. Ullman-Cullere, G. M. Wood, S. M. Lin, M. He, and J. Pathak, "Some experiences and opportunities for big data in translational research," *Genetics in Medicine*, vol. 15, pp. 802-809, Oct 2013.

[10] L. Dai, X. Gao, Y. Guo, J. F. Xiao, and Z. Zhang, "Bioinformatics clouds for big data manipulation," *Biology Direct*, vol. 7, Nov 2012.

[11] V. Marx, "Biology: The big challenges of big data," *Nature,* vol. 498, pp. 255-260, Jun 2013.

[12] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *J Biomed Inform,* vol. 46, pp. 774-781, Oct 2013.

[13] T. B. Murdoch and A. S. Detsky, "The Inevitable Application of Big Data to Health Care," *J Amer Med Assoc,* vol. 309, pp. 1351-1352, Apr 2013.

[14] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Inf Sci Syst,* vol. 2, Feb 2014.

[15] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients," *Health Affairs,* vol. 33, pp. 1123-1131, Jul 2014.

[16] S. Schneeweiss, "Learning from Big Health Care Data," *New Engl J Med,* vol. 370, pp. 2161-2163, Jun 2014.

[17] W. Hsu, M. K. Markey, and M. D. Wang, "Biomedical imaging informatics in the era of precision medicine: progress, challenges, and opportunities," *J Am Med Inform Assn,* vol. 20, pp. 1010-1013, Nov 2013.

[18] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, *et al.*, "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *J Digit Imaging,* vol. 26, pp. 1045-1057, Dec 2013.

[19] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, *et al.*, "Fiji: an open-source platform for biological-image analysis," *Nat Methods,* vol. 9, pp. 676-682, Jul 2012.

[20] H. Banaee, M. U. Ahmed, and A. Loutfi, "Data Mining for Wearable Sensors in Health Monitoring Systems: A Review of Recent Trends and Challenges," *Sensors,* vol. 13, pp. 17472-17500, Dec 2013.

[21] L. D. Xu, W. He, and S. C. Li, "Internet of Things in Industries: A Survey," *IEEE T Ind Inform,* vol. 10, pp. 2233-2243, Nov 2014.

[22] M. Swan, "Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0," *Journal of Sensor and Actuator Networks,* vol. 1, pp. 217-253, Nov 2012.

[23] I. S. Kohane, "Ten things we have to do to achieve precision medicine," *Science,* vol. 349, pp. 37-38, Jul 2015.

[24] F. S. Collins and H. Varmus, "A New Initiative on Precision Medicine," *New Engl J Med,* vol. 372, pp. 793-795, Feb 2015.

[25] R. Margolis, L. Derr, M. Dunn, M. Huerta, J. Larkin, J. Sheehan, *et al.*, "The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data," *J Am Med Inform Assoc,* vol. 21, pp. 957-8, Nov-Dec 2014.

[26] J. Rowley, "The wisdom hierarchy: representations of the DIKW hierarchy," *Journal of Information Science,* vol. 33, pp. 163-180, 2007.

[27] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, *et al.*, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature,* vol. 456, pp. 53-9, Nov 6 2008.

[28] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, *et al.*, "An integrated semiconductor device enabling non-optical genome sequencing," *Nature,* vol. 475, pp. 348-52, Jul 21 2011.

[29] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature,* vol. 437, pp. 376-80, Sep 15 2005.

[30] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, *et al.*, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nature Genetics,* vol. 43, pp. 491-498, May 2011.

[31] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformatics,* vol. 10, Mar 2009.

[32] F. Ozsolak and P. M. Milos, "RNA sequencing: advances, challenges and opportunities," *Nature Reviews Genetics,* vol. 12, pp. 87-98, Feb 2011.

[33] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, *et al.*, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols,* vol. 7, pp. 562-578, Mar 2012.

[34] M. Hirst and M. A. Marra, "Next generation sequencing based approaches to epigenomics," *Brief Funct Genomics,* vol. 9, pp. 455-465, Dec 2010.

[35] S. J. Liu, "Epigenetics advancing personalized nanomedicine in cancer therapy," *Adv Drug Deliver Rev,* vol. 64, pp. 1532-1543, Oct 2012.

[36] S. Pepke, B. Wold, and A. Mortazavi, "Computation for ChIP-seq and RNA-seq studies," *Nature Methods,* vol. 6, pp. S22-S32, Nov 2009.

[37] T. D. Wu and C. K. Watanabe, "GMAP: a genomic mapping and alignment program for mRNA and EST sequences," *Bioinformatics,* vol. 21, pp. 1859-1875, May 2005.

[38] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics,* vol. 25, pp. 1754-1760, Jul 2009.

[39] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics,* vol. 29, pp. 15-21, Jan 2013.

[40] S. Pabinger, A. Dander, M. Fischer, R. Snajder, M. Sperk, M. Efremova, *et al.*, "A survey of tools for variant analysis of next-generation genome sequencing data," *Brief Bioinform,* vol. 15, pp. 256-278, Mar 2014.

[41] H. Li, "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data," *Bioinformatics,* vol. 27, pp. 2987-2993, Nov 2011.

[42] C. Alkan, B. P. Coe, and E. E. Eichler, "Genome structural variation discovery and genotyping," *Nature Reviews Genetics,* vol. 12, pp. 363-375, May 2011.

[43] M. Zhao, Q. G. Wang, Q. Wang, P. L. Jia, and Z. M. Zhao, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives," *BMC Bioinformatics,* vol. 14, Sept 2013.

[44] S. Anders, P. T. Pyl, and W. Huber, "HTSeq—a Python framework to work with high-throughput sequencing data," *Bioinformatics,* vol. 31, pp. 166-169, Jan 2015.

[45] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," *Bioinformatics,* vol. 26, pp. 841-842, Mar 2010.

[46] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome," *Bmc Bioinformatics,* vol. 12, Aug 4 2011.

[47] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, *et al.*, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat Biotechnol,* vol. 28, pp. 511-5, May 2010.

[48] A. McPherson, F. Hormozdiari, A. Zayed, R. Giuliany, G. Ha, M. G. F. Sun, *et al.*, "deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data," *Plos Comput Biol,* vol. 7, May 2011.

[49] D. Kim and S. L. Salzberg, "TopHat-Fusion: an algorithm for discovery of novel fusion transcripts," *Genome Biology,* vol. 12, Aug 2011.

[50] G. Robertson, J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, *et al.*, "De novo assembly and analysis of RNA-seq data," *Nature Methods,* vol. 7, pp. 909-912, Nov 2010.

[51] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, *et al.*, "Full-length transcriptome assembly from RNA-Seq data without a reference genome," *Nature Biotechnology,* vol. 29, pp. 644-U130, Jul 2011.

[52] Z. Chang, G. J. Li, J. T. Liu, Y. Zhang, C. Ashby, D. L. Liu, *et al.*, "Bridger: a new framework for de novo transcriptome assembly using RNA-seq data," *Genome Biology,* vol. 16, Feb 2015.

[53] M. Pertea, G. M. Pertea, C. M. Antonescu, T. C. Chang, J. T. Mendell, and S. L. Salzberg, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nature Biotechnology*, vol. 33, pp. 290-295, Mar 2015.

[54] J. A. Martin and Z. Wang, "Next-generation transcriptome assembly," *Nature Reviews Genetics*, vol. 12, pp. 671-682, Oct 2011.

[55] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, *et al.*, "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs," *Nature Biotechnology*, vol. 28, pp. 503-510, May 2010.

[56] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, *et al.*, "Model-based analysis of ChIP-Seq (MACS)," *Genome Biology*, vol. 9, Sept 2008.

[57] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao, "Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data," *Nucleic Acids Research*, vol. 36, pp. 5221-5231, Sept 2008.

[58] W. S. Bush and J. H. Moore, "Chapter 11: Genome-Wide Association Studies," *Plos Comput Biol*, vol. 8, Dec 2012.

[59] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, pp. 747-753, Oct 2009.

[60] J. R. Gonzalez, L. Armengol, X. Sole, E. Guino, J. M. Mercader, X. Estivill, *et al.*, "SNPassoc: an R package to perform whole genome association studies," *Bioinformatics*, vol. 23, pp. 644-645, Mar 2007.

[61] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly, "A new multipoint method for genome-wide association studies by imputation of genotypes," *Nature Genetics*, vol. 39, pp. 906-913, Jul 2007.

[62] G. T. Wang, B. Peng, and S. M. Leal, "Variant Association Tools for Quality Control and Analysis of Large-Scale Sequence and Genotyping Array Data," *Am J Hum Genet*, vol. 94, pp. 770-783, May 2014.

[63] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, *et al.*, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet,* vol. 81, pp. 559-575, Sept 2007.

[64] J. H. Kim, H. J. Hu, S. H. Yim, J. S. Bae, S. Y. Kim, and Y. J. Chung, "CNVRuler: a copy number variation-based case-control association analysis tool," *Bioinformatics,* vol. 28, pp. 1790-1792, Jul 2012.

[65] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics,* vol. 26, pp. 139-140, Jan 2010.

[66] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology,* vol. 15, Dec 2014.

[67] Y. Hu, Y. Huang, Y. Du, C. F. Orellana, D. Singh, A. R. Johnson, *et al.*, "DiffSplice: the genome-wide detection of differential splicing events with RNA-seq," *Nucleic Acids Research,* vol. 41, Jan 2013.

[68] S. H. Shen, J. W. Park, J. Huang, K. A. Dittmar, Z. X. Lu, Q. Zhou, *et al.*, "MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data," *Nucleic Acids Research,* vol. 40, Apr 2012.

[69] K. Liang and S. Keles, "Detecting differential binding of transcription factors with ChIP-seq," *Bioinformatics,* vol. 28, pp. 121-122, Jan 2012.

[70] H. Xu, C. L. Wei, F. Lin, and W. K. Sung, "An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data," *Bioinformatics,* vol. 24, pp. 2344-2349, Oct 2008.

[71] Y. Zhang, H. B. Liu, J. Lv, X. Xiao, J. Zhu, X. J. Liu, *et al.*, "QDMR: a quantitative method for identification of differentially methylated regions by entropy," *Nucleic Acids Research,* vol. 39, May 2011.

[72] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods,* vol. 8, pp. 469-477, Jun 2011.

[73] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome biology,* vol. 10, 2009.

[74] N. Homer, B. Merriman, and S. F. Nelson, "BFAST: An Alignment Tool for Large Scale Genome Resequencing," *PLoS One,* vol. 4, pp. A95-A106, Nov 11 2009.

[75] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods,* vol. 9, pp. 357-U54, Apr 2012.

[76] Novocraft. *NovoAlign*. Available: http://www.novocraft.com/products/novoalign/

[77] M. David, M. Dzamba, D. Lister, L. Ilie, and M. Brudno, "SHRiMP2: Sensitive yet Practical Short Read Mapping," *Bioinformatics,* vol. 27, pp. 1011-1012, Apr 1 2011.

[78] R. Q. Li, C. Yu, Y. R. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, *et al.*, "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics,* vol. 25, pp. 1966-1967, Aug 1 2009.

[79] Z. M. Ning, A. J. Cox, and J. C. Mullikin, "SSAHA: A fast search method for large DNA databases," *Genome Research,* vol. 11, pp. 1725-1729, Oct 2001.

[80] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology,* vol. 147, pp. 195-7, Mar 25 1981.

[81] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology,* vol. 48, pp. 443-53, Mar 1970.

[82] T. D. Wu and S. Nacu, "Fast and SNP-tolerant detection of complex variants and splicing in short reads," *Bioinformatics,* vol. 26, pp. 873-881, Apr 1 2010.

[83] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics,* vol. 25, pp. 1105-11, May 1 2009.

[84] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, *et al.*, "MapSplice: accurate mapping of RNA-seq reads for splice junction discovery," *Nucleic acids research,* vol. 38, p. e178, Oct 2010.

[85] J. Hu, H. Ge, M. Newman, and K. Liu, "OSA: a fast and accurate alignment tool for RNA-Seq," *Bioinformatics,* vol. 28, pp. 1933-4, Jul 15 2012.

[86] S. Huang, J. Zhang, R. Li, W. Zhang, Z. He, T. W. Lam, *et al.*, "SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data," *Frontiers in genetics,* vol. 2, p. 46, 2011.

[87] H. Li and N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing," *Briefings in Bioinformatics,* vol. 11, pp. 473-483, Sep 2010.

[88] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge, "Analysis and design of RNA sequencing experiments for identifying isoform regulation," *Nature methods,* vol. 7, pp. 1009-15, Dec 2010.

[89] L. Pachter, "Models for transcript quantification from RNA-Seq," *arXiv:1104.3889v2,* 2011.

[90] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature methods,* vol. 5, pp. 621-8, Jul 2008.

[91] S. Lee, C. H. Seo, B. Lim, J. O. Yang, J. Oh, M. Kim, *et al.*, "Accurate quantification of transcriptome from RNA-Seq data by effective length normalization," *Nucleic acids research,* vol. 39, p. e9, Jan 2011.

[92] M. Griffith, O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissy, R. D. Morin, *et al.*, "Alternative expression analysis by RNA sequencing," *Nature Methods,* vol. 7, pp. 843-7, Oct 2010.

[93] R. Bohnert and G. Ratsch, "rQuant.web: a tool for RNA-Seq-based transcript quantitation," *Nucleic acids research,* vol. 38, pp. W348-51, Jul 2010.

[94] J. Feng, W. Li, and T. Jiang, "Inference of isoforms from short sequence reads," *Journal of Computational Biology,* vol. 18, pp. 305-21, Mar 2011.

[95] M. Nicolae, S. Mangul, I. I. Mandoiu, and A. Zelikovsky, "Estimation of alternative splicing isoform frequencies from RNA-Seq data," *Algorithms Mol Biol,* vol. 6, p. 9, 2011.

[96] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinformatics,* vol. 11, p. 94, 2010.

[97] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome biology,* vol. 11, p. R25, 2010.

[98] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome biology,* vol. 11, p. R106, 2010.

[99] M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, *et al.,* "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Brief Bioinform,* Sep 17 2012.

[100] R. K. Patel and M. Jain, "NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data," *Plos One,* vol. 7, Feb 1 2012.

[101] T. H. Stokes, R. A. Moffitt, J. H. Phan, and M. D. Wang, "chip artifact CORRECTion (caCORRECT): a bioinformatics system for quality assurance of genomics and proteomics array data," *Ann Biomed Eng,* vol. 35, pp. 1068-80, Jun 2007.

[102] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, *et al.,* "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nat Rev Genet,* vol. 11, pp. 733-9, Oct 2010.

[103] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics,* vol. 8, pp. 118-27, Jan 2007.

[104] C. Ledergerber and C. Dessimoz, "Base-calling for next-generation sequencing platforms," *Briefings in Bioinformatics,* vol. 12, pp. 489-497, Sep 2011.

[105] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification," *Analytical Chemistry,* vol. 78, pp. 779-787, Feb 1 2006.

[106] S. Simmons, J. Peng, J. Bienkowska, and B. Berger, "Discovering What Dimensionality Reduction Really Tells Us About RNA-Seq Data," *Journal of Computational Biology,* vol. 22, pp. 715-728, Aug 1 2015.

[107] P. L. Auer, S. Srivastava, and R. W. Doerge, "Differential expression-the next generation and beyond," *Briefings in Functional Genomics,* vol. 11, pp. 57-62, Jan 2012.

[108] X. W. Ren, Y. Wang, X. S. Zhang, and Q. Jin, "iPcc: a novel feature extraction method for accurate disease class discovery and prediction," *Nucleic Acids Research,* vol. 41, Aug 2013.

[109] M. Girolami, H. Mischak, and R. Krebs, "Analysis of complex, multidimensional datasets," *Drug Discovery Today: Technologies,* vol. 3, pp. 13-19, Apr 2006.

[110] R. Chen, G. I. Mias, J. Li-Pook-Than, L. Jiang, H. Y. Lam, R. Chen, *et al.*, "Personal omics profiling reveals dynamic molecular and medical phenotypes," *Cell,* vol. 148, pp. 1293-1307, Mar 2012.

[111] L. Stanberry, G. Mias, W. Haynes, R. Higdon, M. Snyder, and E. Kolker, "Integrative Analysis of Longitudinal Metabolomics Data from a Personal Multi-Omics Profile," *Metabolites,* vol. 3, pp. 741-760, Sept 2013.

[112] I. Nookaew, M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlen, *et al.*, "A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae," *Nucleic Acids Research,* vol. 40, pp. 10084-10097, Nov 2012.

[113] N. A. Fonseca, J. Marioni, and A. Brazma, "RNA-Seq Gene Profiling - A Systematic Empirical Comparison," *Plos One,* vol. 9, Sep 30 2014.

[114] P. Y. Wu, J. H. Phan, and M. D. Wang, "Assessing the impact of human genome annotation choice on RNA-seq expression estimates," *BMC Bioinformatics,* vol. 14 Suppl 11, p. S8, 2013.

[115] P. Y. Wu, J. H. Phan, and M. D. Wang, "The Effect of Human Genome Annotation Complexity on RNA-Seq Gene Expression Quantification," *IEEE Int Conf Bioinform Biomed Workshops,* vol. 2012, pp. 712-717, Oct 2012.

[116] P. Y. Wu, J. H. Phan, and M. D. Wang, "An Approach for Assessing RNA-seq Quantification Algorithms in Replication Studies," *IEEE Int Workshop Genomic Signal Process Stat,* vol. 2013, pp. 15-18, Nov 2013.

[117] R. Chandramohan, P. Y. Wu, J. H. Phan, and M. D. Wang, "Benchmarking RNA-Seq Quantification Tools," *2013 35th Annual International Conference of the Ieee Engineering in Medicine and Biology Society (Embc),* pp. 647-650, 2013.

[118] P. Y. Wu, J. H. Phan, F. Zhou, and M. D. Wang, "Evaluation of Normalization Methods for RNA-Seq Gene Expression Estimation," *IEEE Int Conf Bioinform Biomed Workshops,* vol. 2011, pp. 50-57, Nov 2011.

[119] L. C. Bailey, Jr., D. B. Searls, and G. C. Overton, "Analysis of EST-driven gene annotation in human genomic sequence," *Genome Res,* vol. 8, pp. 362-76, Apr 1998.

[120] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, *et al.*, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature,* vol. 447, pp. 799-816, Jun 14 2007.

[121] D. Thierry-Mieg and J. Thierry-Mieg, "AceView: a comprehensive cDNA-supported gene and transcripts annotation," *Genome Biol,* vol. 7 Suppl 1, pp. S12 1-14, 2006.

[122] P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, *et al.*, "Ensembl 2012," *Nucleic Acids Res,* vol. 40, pp. D84-90, Jan 2012.

[123] C. Yamasaki, K. Murakami, Y. Fujii, Y. Sato, E. Harada, J. Takeda, *et al.*, "The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts," *Nucleic Acids Res,* vol. 36, pp. D793-9, Jan 2008.

[124] T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K. O. Koyanagi, *et al.*, "Integrative annotation of 21,037 human genes validated by full-length cDNA clones," *PLoS Biol,* vol. 2, p. e162, Jun 2004.

[125] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Res,* vol. 35, pp. D61-5, Jan 2007.

[126] F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler, "The UCSC Known Genes," *Bioinformatics,* vol. 22, pp. 1036-46, May 1 2006.

[127] L. G. Wilming, J. G. Gilbert, K. Howe, S. Trevanion, T. Hubbard, and J. L. Harrow, "The vertebrate genome annotation (Vega) database," *Nucleic Acids Res,* vol. 36, pp. D753-60, Jan 2008.

[128] G. R. Grant, M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, *et al.*, "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)," *Bioinformatics,* vol. 27, pp. 2518-2528, 2011.

[129] Y. Li, A. Terrell, and J. M. Patel, "Wham: a high-throughput sequence alignment method," in *SIGMOD Conference,* 2011, pp. 445-456.

[130] D. Thierry-Mieg and J. Thierry-Mieg. *Magic Analysis Tool*. Available: ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic/

[131] Y. Liao, G. K. Smyth, and W. Shi, "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote," *Nucleic acids research,* 2013.

[132] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics,* vol. 30, pp. 923-930, Apr 1 2014.

[133] J. W. Bartlett and C. Frost, "Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables," *Ultrasound Obstet Gynecol,* vol. 31, pp. 466-75, Apr 2008.

[134] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychol Bull,* vol. 86, pp. 420-8, Mar 1979.

[135] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nat Rev Genet,* vol. 10, pp. 57-63, Jan 2009.

[136] L. Stein, "Genome annotation: from sequence to biology," *Nat Rev Genet,* vol. 2, pp. 493-503, Jul 2001.

[137] L. Q. Zhang, D. Cheranova, M. Gibson, S. Ding, D. P. Heruth, D. Fang, *et al.*, "RNA-seq reveals novel transcriptome of genes and their isoforms in human pulmonary microvascular endothelial cells treated with thrombin," *PLoS One,* vol. 7, p. e31229, 2012.

[138] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, *et al.*, "Alternative isoform regulation in human tissue transcriptomes," *Nature,* vol. 456, pp. 470-6, Nov 27 2008.

[139] A. Roberts and L. Pachter, "Streaming fragment assignment for real-time analysis of sequencing experiments," *Nature Methods,* vol. 10, pp. 71-U99, Jan 2013.

[140] E. Turro, S. Y. Su, A. Goncalves, L. J. M. Coin, S. Richardson, and A. Lewin, "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads," *Genome Biol,* vol. 12, 2011.

[141] T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, R. Guigo, *et al.*, "Modelling and simulating generic RNA-Seq experiments with the flux simulator," *Nucleic Acids Research,* vol. 40, pp. 10073-10083, Nov 2012.

[142] L. F. Huang, J. M. Jin, P. Deighan, E. Kiner, L. McReynolds, and J. Lieberman, "Efficient and specific gene knockdown by small interfering RNAs produced in bacteria," *Nature Biotechnology,* vol. 31, pp. 350-+, Apr 2013.

[143] S. Pradervand, J. Weber, J. Thomas, M. Bueno, P. Wirapati, K. Lefort, *et al.*, "Impact of normalization on miRNA microarray expression profiling," *RNA,* vol. 15, pp. 493-501, Mar 2009.

[144] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter, "Improving RNA-Seq expression estimates by correcting for fragment bias," *Genome Biol,* vol. 12, p. R22, Mar 16 2011.

[145] L. M. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, *et al.*, "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology,* vol. 24, pp. 1151-1161, Sep 2006.

[146] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, *et al.*, "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nature biotechnology,* vol. 28, pp. 827-838, 2010.

[147] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome research,* vol. 18, pp. 1509-1517, 2008.

[148] S. M.-I. Consortium, "A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium," *Nature Biotechnology,* vol. 32, pp. 903-914, 2014.

[149] R. Lindner and C. C. Friedel, "A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq," *PLoS One,* vol. 7, p. e52403, 2012.

[150] P. G. Engström, T. Steijger, B. Sipos, G. R. Grant, A. Kahles, G. Rätsch, *et al.*, "Systematic evaluation of spliced alignment programs for RNA-seq data," *Nature methods,* vol. 10, pp. 1185-1191, 2013.

[151] A. Hatem, D. Bozdağ, A. E. Toland, and Ü. V. Çatalyürek, "Benchmarking short sequence mapping tools," *BMC bioinformatics,* vol. 14, p. 184, 2013.

[152] I. Borozan, S. N. Watt, and V. Ferretti, "Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq," *PLoS One,* vol. 8, p. e76935, 2013.

[153] A. Kanitz, F. Gypas, A. J. Gruber, A. R. Gruber, G. Martin, and M. Zavolan, "Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data," *Genome biology,* vol. 16, pp. 1-26, 2015.

[154] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nature Biotechnology,* vol. 34, pp. 525-527, May 2016.

[155] E. Maza, P. Frasse, P. Senin, M. Bouzayen, and M. Zouine, "Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: A matter of relative size of studied transcriptomes," *Communicative & Integrative Biology,* vol. 6, 2013.

[156] H. Aanes, C. Winata, L. F. Moen, O. Ostrup, S. Mathavan, P. Collas, *et al.*, "Normalization of RNA-sequencing data from samples with varying mRNA levels," *PloS one,* vol. 9, p. e89158, 2014.

[157] P. Y. Wu, R. Chandramohan, J. H. Phan, W. T. Mahle, J. W. Gaynor, K. O. Maher, *et al.*, "Cardiovascular Transcriptomics and Epigenomics Using Next-Generation Sequencing Challenges, Progress, and Opportunities," *Circulation-Cardiovascular Genetics,* vol. 7, pp. 701-710, Oct 2014.

[158] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC Bioinformatics,* vol. 14, p. 91, 2013.

[159] F. Rapaport, R. Khanin, Y. Liang, M. Pirun, A. Krek, P. Zumbo, *et al.*, "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome biology,* vol. 14, p. R95, Sep 10 2013.

[160] J. Li and R. Tibshirani, "Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data," *Statistical methods in medical research,* vol. 22, pp. 519-36, Oct 2013.

[161] S. Tarazona, F. Garcia-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in RNA-seq: a matter of depth," *Genome research,* vol. 21, pp. 2213-23, Dec 2011.

[162] Z. Chen, J. Liu, H. K. Ng, S. Nadarajah, H. L. Kaufman, J. Y. Yang, *et al.*, "Statistical methods on detecting differentially expressed genes for RNA-seq data," *BMC systems biology,* vol. 5 Suppl 3, p. S1, Dec 23 2011.

[163] L. Wang, Z. Feng, X. Wang, and X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data," *Bioinformatics,* vol. 26, pp. 136-8, Jan 1 2010.

[164] T. J. Hardcastle and K. A. Kelly, "baySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics,* vol. 11, p. 422, 2010.

[165] B. Langmead, K. D. Hansen, and J. T. Leek, "Cloud-scale RNA-sequencing differential expression analysis with Myrna," *Genome biology,* vol. 11, 2010.

[166] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq," *Nature biotechnology,* vol. 31, pp. 46-53, Jan 2013.

[167] E. G. Wilbanks and M. T. Facciotti, "Evaluation of algorithm performance in ChIP-seq peak detection," *PLoS One,* vol. 5, p. e11471, 2010.

[168] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology,* vol. 3, pp. 185-205, 2005.

[169] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation,* vol. 97, pp. 1837-1847, May 12 1998.

[170] G. S. Berenson, S. R. Srinivasan, W. H. Bao, W. P. Newman, R. E. Tracy, W. A. Wattigney, *et al.*, "Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults," *The New England journal of medicine,* vol. 338, pp. 1650-1656, Jun 4 1998.

[171] G. Thanassoulis and R. S. Vasan, "Genetic Cardiovascular Risk Prediction Will We Get There?," *Circulation,* vol. 122, pp. 2323-2334, 2010.

[172] D. K. Arnett, A. E. Baird, R. A. Barkley, C. T. Basson, E. Boerwinkle, S. K. Ganesh, *et al.*, "Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: a scientific statement from the American Heart Association Council on Epidemiology and Prevention, the Stroke Council, and the Functional Genomics and Translational Biology Interdisciplinary Working Group," *Circulation,* vol. 115, pp. 2878-901, Jun 5 2007.

[173] R. B. Schnabel, A. Baccarelli, H. Lin, P. T. Ellinor, and E. J. Benjamin, "Next steps in cardiovascular disease genomic research--sequencing, epigenetics, and transcriptomics," *Clinical chemistry,* vol. 58, pp. 113-26, Jan 2012.

[174] J. S. Ware, A. M. Roberts, and S. A. Cook, "Next generation sequencing for clinical diagnostics and personalised medicine: implications for the next generation cardiologist," *Heart,* vol. 98, pp. 276-281, Feb 2012.

[175] H. K. Song, S. E. Hong, T. Kim, and D. H. Kim, "Deep RNA Sequencing Reveals Novel Cardiac Transcriptomic Signatures for Physiological and Pathological Hypertrophy," *PLoS One,* vol. 7, Apr 16 2012.

[176] W. Guo, S. Schafer, M. L. Greaser, M. H. Radke, M. Liss, T. Govindarajan, *et al.*, "RBM20, a gene for hereditary cardiomyopathy, regulates titin splicing," *Nature medicine,* vol. 18, pp. 766-73, May 2012.

[177] J. Schlesinger, M. Schueler, M. Grunert, J. J. Fischer, Q. Zhang, T. Krueger, *et al.*, "The Cardiac Transcription Network Modulated by Gata4, Mef2a, Nkx2.5, Srf, Histone Modifications, and MicroRNAs," *PLoS Genet,* vol. 7, Feb 2011.

[178] A. He, S. W. Kong, Q. Ma, and W. T. Pu, "Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 108, pp. 5632-7, Apr 5 2011.

[179] M. J. Blow, D. J. McCulley, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, *et al.*, "ChIP-Seq identification of weakly conserved heart enhancers," *Nature genetics,* vol. 42, pp. 806-10, Sep 2010.

[180] C. Mouse Genome Sequencing, R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, *et al.*, "Initial sequencing and comparative analysis of the mouse genome," *Nature,* vol. 420, pp. 520-62, Dec 5 2002.

[181] H. Wu, C. Wang, and Z. Wu, "A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data," *Biostatistics,* vol. 14, pp. 232-43, Apr 2013.

[182] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, "Voom: precision weights unlock linear model analysis tools for RNA-seq read counts," *Genome biology,* vol. 15, p. R29, Feb 3 2014.

[183] P. Delgado-Olguin, Y. Huang, X. Li, D. Christodoulou, C. E. Seidman, J. G. Seidman, *et al.*, "Epigenetic repression of cardiac progenitor gene expression by Ezh2 is required for postnatal cardiac homeostasis," *Nature genetics,* vol. 44, pp. 343-7, Mar 2012.

[184] J. Chen, E. E. Bardes, B. J. Aronow, and A. G. Jegga, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic acids research,* vol. 37, pp. W305-11, Jul 2009.

[185] C. International Human Genome Sequencing, "Finishing the euchromatic sequence of the human genome," *Nature,* vol. 431, pp. 931-45, Oct 21 2004.

[186] A. P. Fejes, G. Robertson, M. Bilenky, R. Varhol, M. Bainbridge, and S. J. Jones, "FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology," *Bioinformatics,* vol. 24, pp. 1729-30, Aug 1 2008.

[187] S. Wilder. (2010). *SWEMBL: a generic peak-calling program*. Available: http://www.ebi.ac.uk/~swilder/SWEMBL/

[188] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng, "A clustering approach for identification of enriched domains from histone modification ChIP-Seq data," *Bioinformatics,* vol. 25, pp. 1952-8, Aug 1 2009.

[189] A. P. Boyle, J. Guinney, G. E. Crawford, and T. S. Furey, "F-Seq: a feature density estimator for high-throughput sequence tags," *Bioinformatics,* vol. 24, pp. 2537-8, Nov 1 2008.

[190] J. T. Robinson, H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, *et al.*, "Integrative genomics viewer," *Nature biotechnology,* vol. 29, pp. 24-6, Jan 2011.

[191] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: scanning for occurrences of a given motif," *Bioinformatics,* vol. 27, pp. 1017-8, Apr 1 2011.

[192] M. Pachkov, I. Erb, N. Molina, and E. van Nimwegen, "SwissRegulon: a database of genome-wide annotations of regulatory sites," *Nucleic acids research,* vol. 35, pp. D127-31, Jan 2007.

[193] D. May, M. J. Blow, T. Kaplan, D. J. McCulley, B. C. Jensen, J. A. Akiyama, *et al.*, "Large-scale discovery of enhancers from human heart tissue," *Nature genetics,* vol. 44, pp. 89-93, Jan 2012.

[194] W. Zhang, Y. Yu, F. Hertwig, J. Thierry-Mieg, W. Zhang, D. Thierry-Mieg, *et al.*, "Comparison of RNA-seq and microarray-based models for clinical endpoint prediction," *Genome biology*, vol. 16, pp. 1-12, 2015.

[195] H. Kitano, "Systems biology: A brief overview," *Science*, vol. 295, pp. 1662-1664, Mar 1 2002.

[196] TCGA. Available: http://cancergenome.nih.gov/

[197] C. D. Kaddi and M. D. Wang, "Models for Predicting Stage in Head and Neck Squamous Cell Carcinoma Using Proteomic Data," *2014 36th Annual International Conference of the Ieee Engineering in Medicine and Biology Society (Embc)*, pp. 5216-5219, 2014.

[198] K. A. Hoadley, C. Yau, D. M. Wolf, A. D. Cherniack, D. Tamborero, S. Ng, *et al.*, "Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin," *Cell*, vol. 158, pp. 929-944.

[199] R. G. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, Y. Qi, M. D. Wilkerson, *et al.*, "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, pp. 98-110, Jan 19 2010.

[200] N. Cancer Genome Atlas, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61-70, Oct 4 2012.

[201] N. Cancer Genome Atlas Research, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, pp. 519-25, Sep 27 2012.

[202] M. K. Keck, Z. Zuo, A. Khattri, T. P. Stricker, C. Brown, M. Imanguli, *et al.*, "Integrative analysis of Head and Neck Cancer identifies two biologically distinct HPV and three non-HPV subtypes," *Clinical Cancer Research*, vol. 21, pp. 870-881, December 9, 2014 2014.

[203] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature Reviews Genetics*, vol. 13, pp. 227-232, Apr 2012.

[204] O. Gottesman, H. Kuivaniemi, G. Tromp, W. A. Faucett, R. Li, T. A. Manolio, *et al.*, "The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future," *Genet Med,* vol. 15, pp. 761-771, Oct 2013.

[205] K. Marsolo and S. A. Spooner, "Clinical genomics in the world of the electronic health record," *Genet Med,* vol. 15, pp. 786-791, Oct 2013.

[206] J. L. Kannry and M. S. Williams, "Integration of genomics into the electronic health record: mapping terra incognita," *Genet Med,* vol. 15, pp. 757-760, Jun 2013.

[207] A. G. Ury, "Storing and interpreting genomic information in widely deployed electronic health record systems," *Genet Med,* vol. 15, pp. 779-785, Aug 2013.

[208] C. A. Caligtan and P. C. Dykes, "Electronic health records and personal health records," *Semin Oncol Nurs,* vol. 27, pp. 218-28, Aug 2011.

[209] G. Alterovitz, J. Warner, P. Zhang, Y. Chen, M. Ullman-Cullere, D. Kreda, *et al.*, "SMART on FHIR Genomics: facilitating standardized clinico-genomic apps," *J Am Med Inform Assoc,* vol. 22, pp. 1173-1178, Nov 2015.

[210] T. D. Schmittgen, B. A. Zakrajsek, A. G. Mills, V. Gorn, M. J. Singer, and M. W. Reed, "Quantitative reverse transcription-polymerase chain reaction to study mRNA decay: comparison of endpoint and real-time methods," *Anal Biochem,* vol. 285, pp. 194-204, Oct 15 2000.

[211] J. Hellemans, G. Mortier, A. De Paepe, F. Speleman, and J. Vandesompele, "qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data," *Genome Biol,* vol. 8, p. R19, 2007.

[212] S. C. Lee, A. Antony, N. Lee, J. Leibow, J. Q. Yang, S. Soviero, *et al.*, "Improved version 2.0 qualitative and quantitative AMPLICOR reverse transcription-PCR tests for hepatitis C virus RNA: calibration to international units, enhanced genotype reactivity, and performance characteristics," *J Clin Microbiol,* vol. 38, pp. 4171-9, Nov 2000.

[213] Y. Karlen, A. McNair, S. Perseguers, C. Mazza, and N. Mermod, "Statistical significance of quantitative PCR," *BMC Bioinformatics,* vol. 8, p. 131, 2007.

VITA

PO-YEN L. WU

WU was born in Taipei, Taiwan. He received a B.S. in Electrical Engineering from National Taiwan University, Taipei, Taiwan in 2008 before joining the graduate school at Georgia Tech. He received a M.S. in Electrical and Computer Engineering in 2011 and a M.S. in Statistics in 2016. He defended his Ph.D. in Electrical and Computer Engineering in 2017. When he is not working on his research, Mr. Wu enjoys spending time with his friends and participating in various indoor and outdoor activities.